Universidade Federal de Santa Catarina
Programa de Pós-Graduação em
Engenharia de Produção

PROPOSIÇÃO DE INDICADORES PARA AVALIAÇÃO TÉCNICA DE PROJETOS DE DATA WAREHOUSE: UM ESTUDO DE CASO NO DATA WAREROUSE DA PLATAFORMA LATTES

Alexandre Marques de Almeida

Dissertação apresentada ao
Programa de Pós-Graduação em
Engenharia de Produção da
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do título de Mestre em
Engenharia de Produção

Florianópolis
2006

Alexandre Marques de Almeida

PROPOSIÇÃO DE INDICADORES PARA AVALIAÇÃO TÉCNICA DE PROJETOS DE DATA WAREHOUSE: UM ESTUDO DE CASO NO DATA WARREROUSE DA PLATAFORMA LATTES

Esta dissertação foi julgada e aprovada para a obtenção do título de **Mestre em Engenharia de Produção** no **Programa de Pós-Graduação em Engenharia de Produção** da Universidade Federal de Santa Catarina

Florianópolis, 10 de março de 2006.

Prof. Edson Pacheco Paladini, Dr.
Coordenador do Curso

BANCA EXAMINADORA

José Leomar Todesco, Dr.
Orientador

Aran Bey Tcholakian Morales, Dr

Vinicius Medina Kern, Dr.

Denilson Sell, Doutorando

Agradecimentos

A Deus por me agraciar com a oportunidade de estar envolvido em um ambiente repleto de pessoas com tamanha capacidade, intelecto e conhecimento auxiliando nesta empreitada tão importante de minha vida.

Agradecimento aos professores José Leomar Todesco - inestimável orientador, Roberto C. S. Pacheco, Aran B. T. Morales e Denílson Sell - co-orientador e amigo, que com seus direcionamentos e sabedoria contribuíram para a conclusão deste trabalho.

Gostaria de agradecer ao Programa de Pós-Graduação em Engenharia de Produção, pela oportunidade em participar de um prestigiado curso. Ao Instituto Stela, onde foi possível a realização deste estudo.

Gostaria de agradecer também aos amigos do Instituo Stela que de alguma forma ajudaram.

Aos meus pais, Adicionil e Zenaide, e familiares por sempre me apoiarem e incentivarem incondicionalmente.

E finalmente, agradecer à Alessandra Acácia Alves pelo seu companheirismo.

"O principal objetivo de todo o progresso técnico deve ser o homem e o seu destino...para que as criações da nossa inteligência possam ser uma bênção, e nunca uma maldição para a humanidade"

Albert Einstein

Resumo

Como peça fundamental na obtenção do conhecimento, a informação requer cada vez mais o uso de tecnologias de computação. Esta necessidade se traduz na aplicação de ferramentas que possam agregar grandes massas de dados armazenadas ao longo do tempo como base de conhecimento e transformá-las em indicadores sustentáveis para futuras tomadas de decisões. Com a utilização de técnicas de *data warehousing* é possível obter tais indicadores para tomadas de decisões. Porém, no próprio processo de *data warehousing* existe dificuldade em se obter indicadores de desenvolvimento, seja na modelagem dos dados, no projeto de *back-end* ou de *front-end* de um *DW*. Esta dissertação propõem com o estudo de caso no *data warehouse* da Plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a utilização de indicadores relacionados a modelagem de dados, o projeto de *back-end* e *front-end* no processo de *data warehousing*, encontrados nos *data marts* de Fomento, Grupos de Pesquisa e *DM* de Currículos que são integrantes do *DW* da Plataforma Lattes e que apresentam modelos diferenciados possibilitando a aplicação dos indicadores e verificação da atuação de cada um dos indicadores em cada modelo. O presente trabalho visa auxiliar no desenvolvimento de novos projetos de *DW*, e diminuir a carência de pesquisas realizadas sobre o levantamento e utilização de indicadores de desenvolvimento de *DW*.

Abstract

As an essential piece to obtain knowledge, information needs computer technology more and more. To fill this need, tools are applied in order to aggregate a big mass of data that has been stored for a long time as a knowledge base. Then the data is transformed in tenable indicators to be used in the decision making process. Through data warehousing techniques, it's possible to obtain these indicators. However, in the data warehousing process, it's hard to obtain development indicators when implementing the data model, the back-end or the front-end project. This thesis presents a use case of a DW implementation for a project named Lattes Platform, from the Scientific and Technological Development National Counsel (CNPq). This project uses indicators related to the data model, back-end and front-end projects from the data warehousing process, found in several data marts: Fomentation, Research Groups and Curriculum. Those data marts compose the Lattes Platform DW and present differentiated models that enable the application of indicators and verification of these indicators action in each model. The present work aims at assisting the development of new DW projects in order to reduce research deficiency related to indicators gathering and use for the DW implementation.

Sumário

RESUMO ..... **V**

ABSTRACT ..... **VI**

LISTA DE FIGURAS ..... **IX**

LISTA DE TABELAS ..... **X**

LISTA DE REDUÇÕES ..... **XI**

1 INTRODUÇÃO **1**

1.1 INTRODUÇÃO 1
1.2 OBJETIVO 2
1.2.1 OBJETIVOS ESPECÍFICOS 3
1.3 JUSTIFICATIVA 3
1.4 METODOLOGIA 4
1.5 ESTRUTURA DO TRABALHO 5

2 DATA WAREHOUSE **6**

2.1 INTRODUÇÃO 6
2.2 DEFINIÇÃO DE *DATA WAREHOUSE* 6
2.3 ARQUITETURAS DE *DATA WAREHOUSE* 7
2.3.1 ARQUITETURA *TOP-DOWN* 7
2.3.2 ARQUITETURA *BOTTOM-UP* 9
2.3.3 ARQUITETURA HÍBRIDA 10
2.3.4 ARQUITETURA FEDERADA 12
2.4 ESTRUTURA DO REPOSITÓRIO DO *DATA WAREHOUSE* 13
2.4.1 DIMENSIONAL 14
2.4.2 NORMALIZADO 15
2.4.3 SNOWFLAKE 16
2.5 *BACK-END* 17
2.5.1 PROJETO FÍSICO 18
2.5.2 PROCESSO DE *ETL* 20
2.5.3 ESTRATÉGIAS DE OTIMIZAÇÃO DE REPOSITÓRIO 23
2.6 *FRONT-END* 25
2.7 CONCLUSÃO DO CAPÍTULO 26

3 O *DATA WAREHOUSE* DO CNPQ **27**

3.1 INTRODUÇÃO 27
3.2 A ARQUITETURA DO *DATA WAREHOUSE* DO CNPQ 31
3.3 O *DATA MART* DE FOMENTO 32
3.3.1 MODELAGEM 33
3.3.2 *BACK-END* 34
3.3.3 *FRONT-END* 35
3.4 O *DATA MART* DE GRUPOS DE PESQUISA 37
3.4.1 MODELAGEM 37
3.4.2 *BACK-END* 39
3.4.3 *FRONT-END* 39
3.5 O *DATA MART* DE CURRÍCULOS 41
3.5.1 MODELAGEM 41
3.5.2 *BACK-END* 43
3.5.3 *FRONT-END* 43
3.6 CONCLUSÃO DO CAPÍTULO 45

4 O ESTUDO REALIZADO **46**

4.1 INTRODUÇÃO 46
4.2 MODELAGEM 47

4.2.1 Redundância de informação 48
4.2.2 Volatilidade dos dados 49
4.2.3 Temporalidade 49
4.2.4 Flexibilidade de ajuste 50
4.2.5 Granularidade 51
4.2.6 Agregados 52
4.2.7 Reutilização de modelos 53
4.3 BACK-END 53
4.3.1 Processo de carga (Performance, eficiência) 54
4.3.2 Volume de dados e índices 55
4.3.3 Indexação 56
4.3.4 Criação de agregados 57
4.4 *FRONT-END* 58
4.4.1 Legibilidade 59
4.4.2 Utilização de OLAP 59
4.4.3 Consultas (Performance, complexidade) 60
4.5 RESULTADOS 62
4.6 CONCLUSÃO DO CAPÍTULO 64
5 CONCLUSÕES E TRABALHOS FUTUROS 65
REFERÊNCIAS BIBLIOGRÁFICAS 67
ANEXO I – MODELO DO *DM* DE FOMENTO 70
ANEXO II – MODELO DO *DM* DE GRUPOS DE PESQUISA 71
ANEXO III – MODELO DO *DM* DE CURRÍCULOS 72

Lista de Figuras

Figura 1 – Arquitetura *Top-Down* 8
Figura 2 – Arquitetura *Bottom-Up* 9
Figura 3 – Arquitetura Híbrida 11
Figura 4 – Arquitetura Federado 12
Figura 5 – Representação do modelo dimensional como um cubo de dados 14
Figura 6 – Modelo dimensional do *DW* (estrela) 15
Figura 7 – Modelo *Snowflake* 17
Figura 8 – Processo de *Back-End* 18
Figura 9 – Arquitetura conceitual para projetos de E-Gov. 29
Figura 10 – Processo de carga e publicação dos dados 32
Figura 11 – Modelo estrela simplificado do *Data Mart* de Fomento 34
Figura 12– Site do *Data Mart* de Fomento 36
Figura 13 - Site do *Data Mart* de Fomento 36
Figura 14 – Modelo parcial do *DM* de Grupo de Pesquisa 39
Figura 15 – Séries históricas de grupos por região 40
Figura 16 – Plano tabular de grupos de pesquisa por área de conhecimento 40
Figura 17 – Modelo parcial do DM de Currículo 42
Figura 18 – Demografia Curricular 44
Figura 19 – Lattes Egressos 45

Lista de Tabelas

Tabela 1 – Dimensões e Fatos do *DM* de Fomento 33
Tabela 2 - Dimensões e Fatos do *DM* de Grupos de Pesquisa 37
Tabela 3 - Dimensões e Fatos do *DM* de Currículos 41
Tabela 4 - Indicadores Comparativos 62

Lista de Reduções

Siglas

C&T Ciência e Tecnologia
CAPES Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
CTC Conselho Técnico-Científico
DM *Data Mart*
DW *Data Warehouse*
ETL *Extraction, Transformation and Loading*
(Extração, Transformação e Carga)
ODS *Operational Data Store*
OLAP *On-Line Analytical Processing* (Processamento Analítico *On-Line*)
OLTP *On-Line Transaction Processing* (Processamento de Transações *On-Line*)
TIC Tecnologias de Informação e da Comunicação
XML Extensible Markup Language (linguagem de marcação de dados)

1 INTRODUÇÃO

1.1 INTRODUÇÃO

Cada vez mais a informação vem se tornando peça fundamental na obtenção do conhecimento, e esta informação sendo gerada de forma coesa e de qualidade auxilia as decisões a serem tomados pelos gestores administrativos das empresas. Não diferentes são as maneiras de como o governo também toma suas decisões na atualidade, a fim de proporcionar serviços melhorados, informação confiável e conhecimento para facilitar o processo de governo e encorajar a participação do cidadão. Para que se chegue a esses objetivos, tecnologias de informação e da comunicação (TIC) estão sendo amplamente utilizadas e, dentre elas, nas produções de sistemas de informações, a técnica de *data warehousing* ou criação de *Data Warehouse* (*DW*).

Com a utilização massiva de *data warehousing* nas organizações, problemas no seu desenvolvimento e conclusão surgiram e tendem a prejudicar no seu sucesso. Os usuários podem oferecer resistência a sua utilização, ou um erro de definição e escolha de requisitos ou modelos, tornam-se indicadores de possíveis fracassos. Para evitar tal desfecho é necessário fontes de conhecimento embasadas em casos bem sucedidos, ou em caso de projetos não concluídos, e talvez concluídos mas abandonados, que demonstrem em que momentos de todo o processo houve engano na escolha ou definição a ser seguida, diminuindo assim as possibilidades de fracasso em futuros projetos de *DW*.

Devido a carência de estudos mais aprofundados que apontem indicadores que demonstrem qual o modelo mais adequado ao assunto que será foco da criação de um *DW*, torna-se necessário a análise de casos reais a fim de obter critérios que forneçam maior segurança e eficácia ao resultado final do projeto de *DW.* Este trabalho realiza essa análise através do estudo comparativo entre três modelos de *data warehousing* onde levantou-se indicadores característicos de relevância na construção dos mesmos e que podem ser utilizados como base de conhecimento em outros projetos. O

indicadores avaliados estão divididos nos processos de modelagem, *back-end* e *front-end* de construção do *DW* onde na modelagem avalia-se a redundância de informação, volatilidade dos dados, temporalidade da informação, flexibilidade de ajustes no modelo, granularidade da informação, nível de agregação da informação e a reutilização do modelo em novos projetos. No processo de *back-end* os indicadores avaliados são o processo de carga dos dados do *DW*, o volume dos dados e índices, os métodos de indexação e a criação de agregados. Finalmente no processo de front-end forma avaliados indicadores de legibilidade da informação, a possibilidade de utilização de ferramentas *OLAP* sobre as informações armazenadas no *DW* e a performance de resposta as consultas realizadas pelos usuários.

Para a realização do estudo foram usados como exemplo os *Data Marts* integrantes do *DW* da Plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que consiste na reunião e organização dos dados em *Data Marts* (*DM*) gerados por um conjunto de sistemas computacionais denominados Plataforma Lattes que compatibiliza e integra as informações em toda interação da Agência com seus usuários. A criação desses repositórios de dados gera informação de suma relevância sobre a atividade científica e tecnológica do Brasil, assim como possibilita importantes análises realizadas pelos técnicos do CNPq que tornam mais ágeis e transparente as ações da Agência e disponibilizam o acesso mais rápido a essas informações pelos grupos de pesquisa, pesquisadores e bolsistas do país.

Os *DMs* analisados foram o *Data Mart* de Fomento que foi concebido sobre um modelo dimensional analítico de *DW*, *Data Mart* de Grupos de Pesquisa que também é um modelo dimensional analítico mas com a utilização de novas dimensões denominadas “dimensões pontes” e *Data Mart* de Currículos do CNPq que é um modelo totalmente agregado caracterizado pelos inúmeros fatos que atendem os assuntos determinados no levantamento de requisitos do processo de *data warehousing*.

1.2 OBJETIVO

Este trabalho tem como objetivo avaliar as diferentes abordagens empregados no processo de desenvolvimento de uma arquitetura de *DW*, como a construção de

modelos puramente dimensionais ou modelos totalmente baseados em agregados, levando em consideração as características pertinentes ao assunto que será abordado como tema do *data warehouse* ou até mesmo de cada um dos *data marts* relacionados, para auxiliar na escolha do modelo de dados a ser utilizado em outros projetos de *DW*.

1.2.1 OBJETIVOS ESPECÍFICOS

- elaborar conjunto de indicadores para avaliação que possam ajudar na identificação de vantagens na abordagem adotada em cada um dos *Data Marts*;
- analisar os modelos dos *Data Marts* integrantes do *DW* da Plataforma Lattes que são o *Data Mart* de Fomento, *Data Mart* de Grupos de Pesquisa e *Data Mart* de Currículos do CNPq a partir da lista de indicadores identificados;

1.3 JUSTIFICATIVA

De acordo com Sinha & Sen [2005], o amadurecimento do mercado tende a convergir as metodologias de data warehousing, mas ainda não existem metodologias que sejam reconhecidas como um padrão utilizado por todos. Existem metodologias de vendedores que podem ser utilizadas, mas em organizações que tem total conhecimento de seu negócio, podendo criar então modelos de informação.

Este mercado de soluções de *DW* vale mais de 7 bilhões de dólares mundialmente e pode dobrar até 2006 [COMPUTERWORLD, 2004] e de acordo com a Forrester Research, mais de 44% das companhias irão adotar soluções de *DW* durante o ano de 2005 [BUSINESSINTELLIGENCE.COM, 2004]. Apesar deste cenário propício o Gartner Group [2004] com a ComputerWorld [2004], prevêem que até 2007, cerca de 50% dos projetos de *DW* terão problemas de aceitação por seus usuários ou acabarão por ser mal sucedidos em seus objetivos finais.

Determinar a melhor escolha da arquitetura no desenvolvimento de um *DW* implica diretamente no seu sucesso, e para isso uma visão clara no levantamento de requisitos deve ser baseada em prévios conhecimentos que podem ser direcionados por indicadores de relevância em *data warehousing.* Um estudo comparativo entre projetos

de *DW* existentes e em plena produção pode apresentar tais indicadores e possibilitar melhor tomada de decisões na construção de futuros *DWs*.

A necessidade de abordar esse tema não se justifica apenas pelas deficiências identificadas e pela importância dessas ferramentas no processo decisório das organizações, mas também pela carência de pesquisas que abordem especificamente esse problema.

1.4 METODOLOGIA

Para efetivar os objetivos, o trabalho fundamenta-se em três etapas:

1) estudo sobre a teoria acerca de *data warehousing* e das principais metodologias empregadas nesta tecnologia;
2) definir indicadores de relevância a concepção de modelos de dados no processo de desenvolvimento de *data warehouse*;
3) estudo analítico sobre os modelos de dados, processo de *back-end* e *front-end* dos *data marts* do CNPq.

Na primeira etapa são levantadas as características teóricas na construção de um *data warehouse* onde são explanadas as arquiteturas mais utilizadas, as estruturas típicas de repositórios de dados, os processos de *ETL* e otimização do acesso aos dados e as ferramentas *OLAP* utilizadas pelos usuários dos dados.

Na segunda etapa são apresentados os *DM* do CNPq, seus modelos de dados, os processos de extração dos dados origem para o *DM*, limpeza e transformação desses dados, carga dos dados no repositório final, processos de otimização de consultas aos dados e ferramentas de apresentação das informações contidas no *DM* ao usuário final.

Na terceira etapa são identificados os indicadores que se tornam de maior relevância em cada parte do processo de *data warehouseing* e que afetarão no sucesso da conclusão de um *DW*.

1.5 ESTRUTURA DO TRABALHO

A estrutura do trabalho compreende os seguintes capítulos:

- Capítulo 2: *Data Warehouse* – trata o conteúdo teórico na construção de *DWs* apresentando as arquiteturas de desenvolvimento mais utilizadas, a estruturação dos repositórios de dados com seus tipos de modelos, o projeto físico onde se determina o sistema gerenciador de banco de dados (SGBD) o modelo físico e se estima o volume de dados que será inserido no repositório de *DW* e seu crescimento, indexação e particionamento de arquivos de dados, estratégias de otimização no acesso aos dados. Também são explanados o *back-end* com as origens dos dados para formar o repositório, os sistemas de extração, transformação e carga de dados (*ETL*) e o *front-end* que é a apresentação das informações aos usuários finais;
- Capítulo 3: O *Data Warehouse* do CNPq – é abordado o *DW* do CNPq, onde são vistos os *Data Marts* de Fomento, Grupos de Pesquisa e de Currículos, que fazem parte do objeto de estudo dos modelos de dados realizado neste trabalho;
- Capítulo 4: O Estudo Realizado – este capítulo relata os aspectos mais detalhados dos modelos e identifica os indicadores de maior relevância a concepção de cada modelo, na busca de auxiliar futuros trabalhos de desenvolvimento de *DWs*;
- Capítulo 5: Conclusões e Recomendações – são relatadas as conclusões do trabalho, assim como recomendações e projetos para trabalhos futuros.

2 DATA WAREHOUSE

2.1 INTRODUÇÃO

Este capítulo descreve o ambiente de um *data warehouse* e suas arquiteturas, formas de estruturação dos repositórios de dados, o levantamento do projeto físico, as estratégias de otimização dos repositórios de dados, todo o processo de extração, limpeza e carga dos dados de um repositório operacional para um repositório de *data warehouse* (*back-end*), assim como os processos necessários para a interação do usuário final com o *data warehouse* (*front-end*).

A apresentação do processo de *data warehousing* é necessária para realizar o levantamento de indicadores dentro do processo de modelagem de dados, projeto de *back-end* e *front-end* que serão utilizados como base de conhecimento no projeto de novos *DW*s.

2.2 DEFINIÇÃO DE *DATA WAREHOUSE*

Segundo Singh [2001], *DW* é uma tecnologia de gestão e análise de dados, constituindo "um ambiente de suporte a decisão que alavanca dados armazenados em diferentes fontes e os organiza e entrega aos tomadores de decisões da empresa, independente de plataforma que utilizam ou de seu nível de qualificação técnica".

Já Inmon [1997] diz que *data warehouse* é "um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais".

Um ambiente de *DW* pode ser concebido a partir de dados operacionais oriundos de sistemas *ODS* (*Operational Data Store*) legados ou de dados externos e também de maneira lógica onde seria a união de *Data Marts* (*DM*) que são conjuntos flexíveis de

dados orientados por cada assunto estratégico de negócio e que também tem como origem de dados sistemas *ODS* ou dados externos.

2.3 ARQUITETURAS DE *DATA WAREHOUSE*

Cada *data warehouse* é distinto, isso porque deve ser moldado de forma a atender as necessidades dos usuários de negócio em suas áreas funcionais dentro de uma empresa cujas condições de negócio e as pressões de competitividade são diferentes. Contudo quatro arquiteturas são as mais utilizadas no desenvolvimento e comumente chamadas de:

- Arquitetura *top-down*;
- Arquitetura *bottom-up*;
- Arquitetura híbrida;
- Arquitetura federada.

O plano de arquitetura é um fator importante na construção do projeto de um *DW* como ferramenta de comunicação, planejamento e flexibilidade, que facilita o aprendizado e aumento da produtividade [Kimball, 1998].

Contudo, para um melhor entendimento, a explanação das arquiteturas mais utilizadas torna-se necessária para se obter uma visão clara das sutilezas entre cada abordagem auxiliando no rumo a ser tomado para o sucesso na obtenção de um *data warehouse* que atenda as necessidades de informação da organização.

2.3.1 ARQUITETURA *TOP-DOWN*

Defendida por Inmon [1997], a abordagem *top-down* visualiza o *Data Warehouse* como o centro do ambiente analítico inteiro da empresa. O *DW* possui atomicidade ou transação de dados que são extraídos de um ou mais sistemas e integrados dentro de uma modelagem de dados normalizada da empresa. Neste ambiente os dados são resumidos, dimensionados, e distribuídos para um ou mais *Data Marts* dependentes como mostra a Figura 1. Estes *DM* são “dependentes” porque eles

derivam todos os seus dados de um *DW* centralizado. Às vezes, as organizações suprem os *DWs* com uma área de estagiamento para buscar dados de sistemas fontes antes de poderem ser movidos e integrados dentro do *DW*. Uma área de estagiamento separada é particularmente útil se existem numerosos sistemas fonte ou grandes volumes de dados a serem processados e integrados ao *DW*.

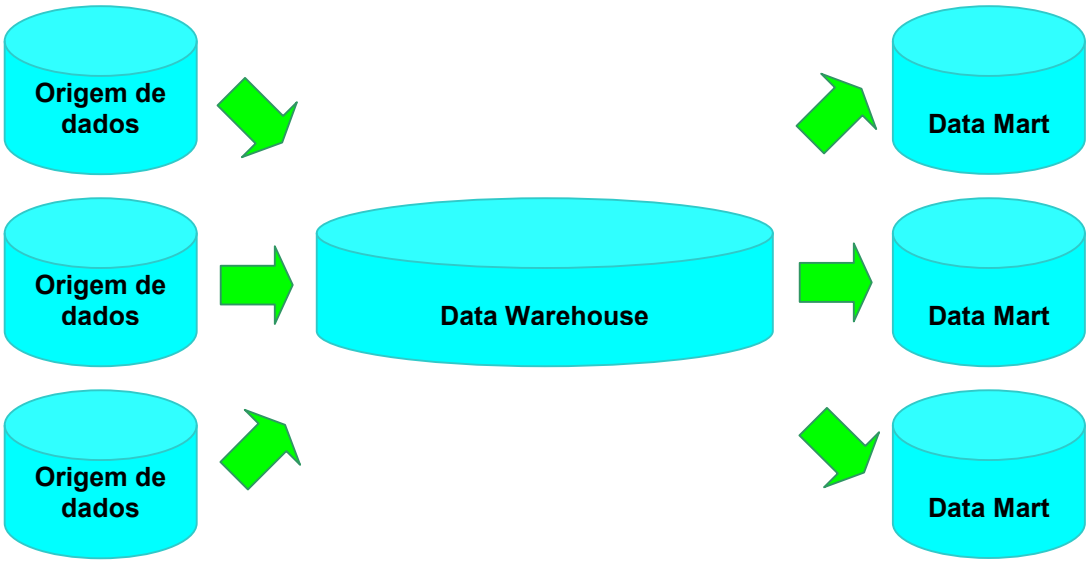


Figura 1 – Arquitetura *Top-Down*

Como vantagens dessa abordagem pode-se citar como mais importante a criação de uma arquitetura integrada, flexível capaz de suportar a inclusão de novos assuntos de interesse para análise da organização em suas estruturas de dados. Isto significa que um *DW* é o ponto de partida para a criação de todos os *DM* obrigando uma consistência e padronização de forma que a organização possa alcançar uma única visão sobre a informação. Os dados atômicos no *DW* permitem a organização reavaliar os dados por vários ângulos para encontrar novas e inesperadas necessidades de negócio. Um *DW* pode ser usado para criar informações para estatísticas, emitir relatórios operacionais, ou dar suporte a *ODSs* e aplicações analíticas. Além disso os usuários podem consultar o *DW* se precisarem cruzar dados ou visões de empreendimento.

Por outro lado, o uso da abordagem *Top-Down* é de desenvolvimento a longo prazo e tende a custar mais caro a organização, principalmente na fase inicial. Isto porque a organização deve criar um modelo de dados razoavelmente detalhado da empresa assim como disponibilizar a infra-estrutura física necessária para alojar a área de estagiamento, o *DW* e os *DM* antes de desenvolver suas aplicações e relatórios. Essa

demora inicial pode provocar o desenvolvimento de aplicações analíticas por grupos isolados da organização que tem independência orçamentária, gerando um ambiente heterogêneo onde sistemas de análise não buscariam seus dados de uma única fonte.

2.3.2 ARQUITETURA *BOTTOM-UP*

Esta abordagem é apresentada por Kimball [1998] e tem em sua característica principal um processo de desenvolvimento inverso a proposta de Immon. Primeiramente são criados os *Data Marts* a partir de dados de sistemas legados ou dados externos podendo assim gerar um *DW* ou apenas considerar como *DW* a união de todos os *DM*, como representado na Figura 2.

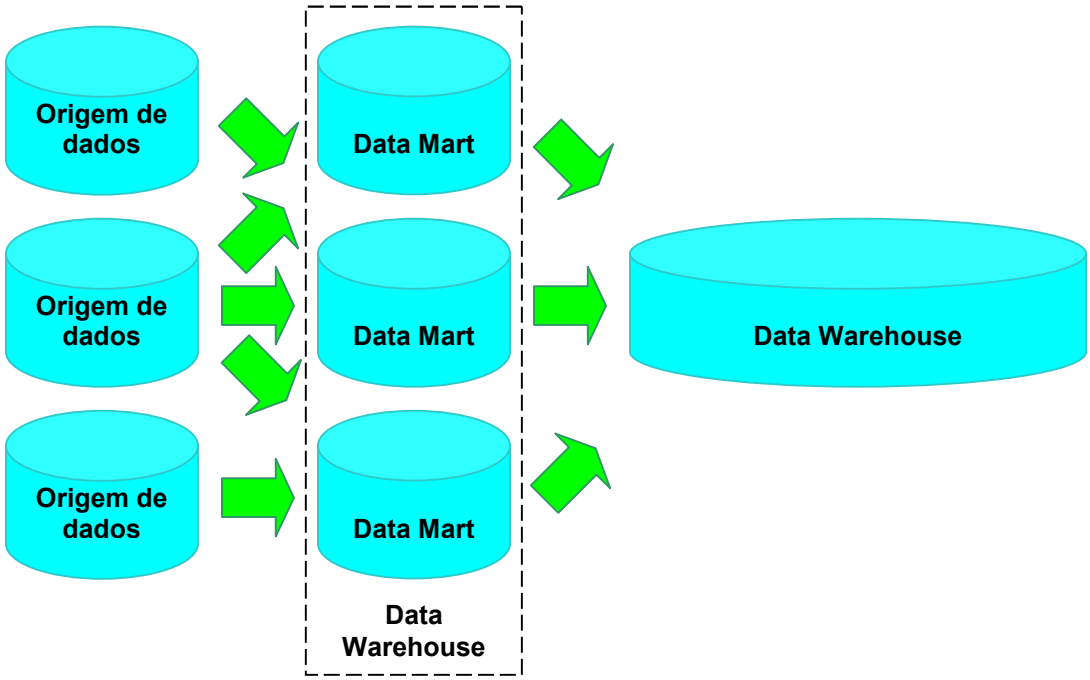


Figura 2 – Arquitetura *Bottom-Up*

Nesta abordagem a meta é desenvolver os *DM* o mais rápido possível apresentando os dados juntos de maneira tanto atômica quanto resumida que os usuários possam querer agora ou no futuro. Os *DM* são desenvolvidos focando um assunto da organização por *DM* e sempre observando a reutilização de dimensões ou fatos conformados, ou seja, que podem ser comum a todos. Estes *DMs* são criados

levando em consideração a integração entre cada *DM* para se obter uma visão única de todo o negócio formando assim um *DW.*

Tornas-se vantajosa essa abordagem pois são desenvolvidos primeiros os *DMs* sobre os assuntos de maior interesse da organização, não despendendo esforços e recursos em *DMs* de assuntos de pouco interesse inicialmente. Posteriormente o ambiente fornece maior facilidade na integração de novos *DMs* de assuntos que forem surgindo. Isto reflete também na infra-estrutura necessária para abrigar todo o *DW* que nesta abordagem o seu custo se torna bem menor. Com o rápido desenvolvimento as respostas aos usuários conseqüentemente são também rápidas gerando um clima de confiança e satisfação em todo o projeto.

Um problema nesta abordagem vem do fato da obrigatoriedade da organização em forçar o uso das dimensões e fatos conformados para que se tenha uma visão única de todo o negócio na organização. Isso em organizações distribuídas e descentralizadas pode vir a acontecer pois seus departamentos e unidades de negócio tendem a reutilizar suas próprias referências e regras para gerar seus fatos e formar *Data Marts* independentes ou não integrados.

Além disso, *DMs* são desenhados para otimizar consultas e não para abrigar execuções em lote ou muitos processos transacionais. Organizações que usam a abordagem *Bottom-Up* necessitam criar uma estrutura fora dessa arquitetura para utilizar requisições de *Data Mining, ODS,* e relatórios operacionais. Porém isso pode ser amenizado buscando um subconjunto de dados do *DM* necessário para tais operações durante a noite quando os usuários não estão utilizando os sistemas.

2.3.3 ARQUITETURA HÍBRIDA

Segundo Hackney [1998], esta estratégia tem a finalidade de integrar a estratégia "*Top-Down*" com a "*Bottom-Up*". Essa abordagem tende a utilizar o que existe de melhor na duas estratégias, tenta tirar proveito da velocidade e orientação do usuário existente na estratégia "*Bottom-Up*" sem sacrificar a integração forçada por um *Data Warehouse* na estratégia "*Top-Down*". De início é modelado todo o *Data Warehouse* da organização sendo então assim implementado partes desse modelo que

representam os assuntos da organização e que se tornarão os *Data Marts*. Como mostra a figura 3.

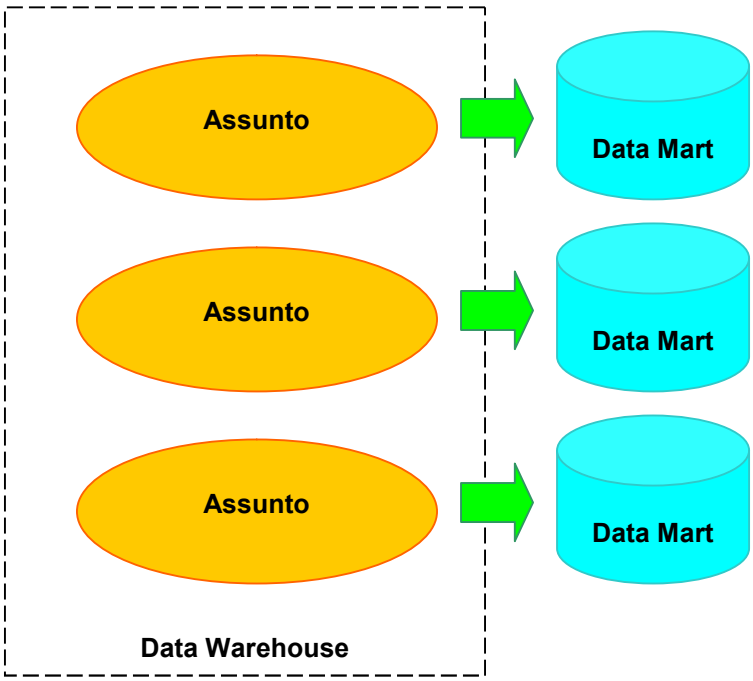


Figura 3 – Arquitetura Híbrida

Esta abordagem conta com ferramentas *ETL* para armazenar e gerenciar os modelos locais e da empresa em *Data Marts* assim como sincronizar as diferenças entre eles, deixando assim que os departamentos desenvolvam suas próprias definições e regras para os elementos de dados que são derivados a partir do modelo da empresa sem sacrificar a integração a longo prazo. Tais ferramentas também são usadas para extrair e carregar dados de sistemas fontes dentro dos *Data Marts* em ambos os níveis, atômico e sumarizado. A organização então transfere os dados atômicos dos *DMs* para o *DW* e alimenta os fatos consolidados, economizando tempo, dinheiro e recursos de processamento.

O maior benefício dessa abordagem é a combinação das técnicas mais rápidas de desenvolvimento onde é implementado um modelo de dados interativo da empresa e apenas desenvolvido as estruturas de real importância e necessárias. Porém depende fortemente de uma ferramenta *ETL* para sincronizar os meta dados entre a empresa e departamentos distribuídos, gerar agregados, carregar dados detalhados e gerenciar a transição para a infra-estrutura do *DW*.

2.3.4 ARQUITETURA FEDERADA

Defendida por Hackney [1998], a abordagem federado não é uma metodologia ou uma arquitetura por si só, mas uma concessão para a falta de apoio que enfraquecem os melhores planos para desenvolver um sistema perfeito. O *Data Warehouse* federado provê uma interface que o torna semelhante um grande *DW*, mas que na realidade, apenas adiciona camadas sobre os *DW*s existentes, possibilitando a execução de consultas sobre um grande *Data Warehouse* "virtual" (HUBER et al., 2001) .

Basicamente a formação de um *DW* federado é composta por uma camada de *DWs* existentes, heterogêneos e distribuídos que são acessados por uma camada de integração que organiza os dados de forma homogênea e carrega o *DW* federado, como mostra a Figura 4.

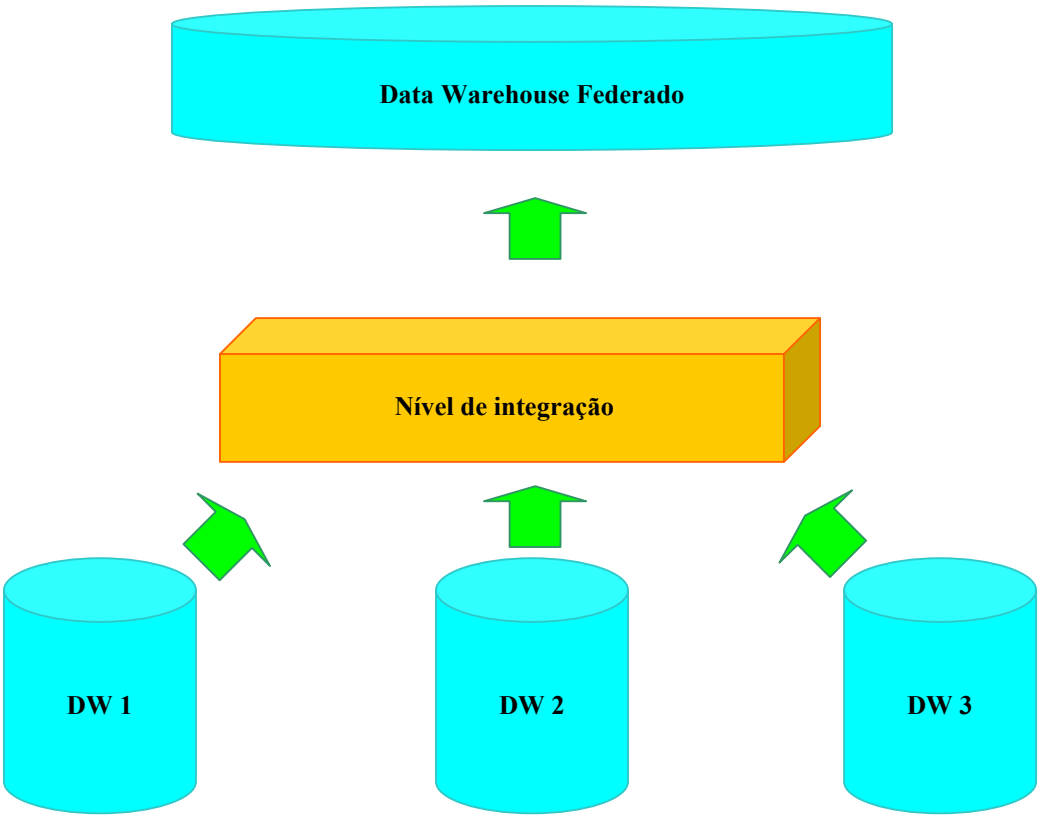


Figura 4 – Arquitetura Federado

Uma definição importante nesta abordagem é como será constituída a camada do *DW* federado. Está pode ser apenas criada com metadados que são gerados pela camada de integração e que auxiliam as consultas requeridas pela camada mais alta a

buscarem os dados nos *DWs* distribuídos e retornam diretamente o resultado ao requerente. Outra forma é materializar a camada do *DW* federado, de maneira que a camada de integração busque os dados nos *DW*s distribuídos, tratem esses dados e os depositem em um repositório físico.

Levando em consideração a materialização da camada do *DW* federado, isso tornar a respostas as consultas muito mais rápidas principalmente quando se é necessário informação detalhada de vários *DWs.* Além de um melhor resultado no processo de *ETL* onde pode-se consistir melhor os dados antes de depositá-los no repositório do *DW* federado, a infra-estrutura requerida também é menor devido a pouca necessidade de alto poder de processamento das consultas. Já a opção pela criação de metadados na camada do *DW* federado evita todo o processo de *ETL* e torna muito mais rápido para se disponibilizar novos dados ou a integração de novos *DWs* ao conjunto, mas depende de uma infra-estrutura poderosa capaz de suportar o alto processamento realizado para se obter os resultados das consultas requeridas.

2.4 ESTRUTURA DO REPOSITÓRIO DO *DATA WAREHOUSE*

A construção do modelo de dados é fundamental para o desenvolvimento, ajudando a compreender as regras de negócio e a organização dos dados para o melhor tempo de resposta, Kimball [1998].

O modelo de dados do *DW* representa os requisitos de informações integradas, os requisitos de análise e o suporte à decisão de toda a organização, Singh [2001].

A representação do modelo de dados de um *DW* é mais simples que os modelos Entidade/Relacionamento utilizados nos sistemas de processamento de transações *on-line* (*OLTP),* pois estes modelos são constituídos de poucas tabelas de dimensões ligadas a uma ou mais tabelas de fatos. Já os modelos destinados a sistemas *OLTP* são um emaranhado de relacionamentos com um número grande de tabelas e quase sempre normalizados. Estes modelos não podem ser utilizados para um *DW* pois são dificilmente entendidos pelos usuários e tem uma performance baixa em consultas direcionadas a informações geradas por um *DW*. Estes modelos serão abordados em suas características a seguir.

2.4.1 DIMENSIONAL

No ambiente de *DW*, é utilizado o modelo dimensional que utiliza uma técnica que suporta o ambiente para análise multidimensional dos dados, Machado [2000].

A representação do negócio de uma organização pode ser feita através de um "cubo de dados", onde cada ponto interno contém as medidas relativas à combinação das dimensões do dado, como mostra a Figura 5, sendo as vendas as medições numéricas do negócio (fato) e tempo, produto e local as medidas do negócio (dimensões).

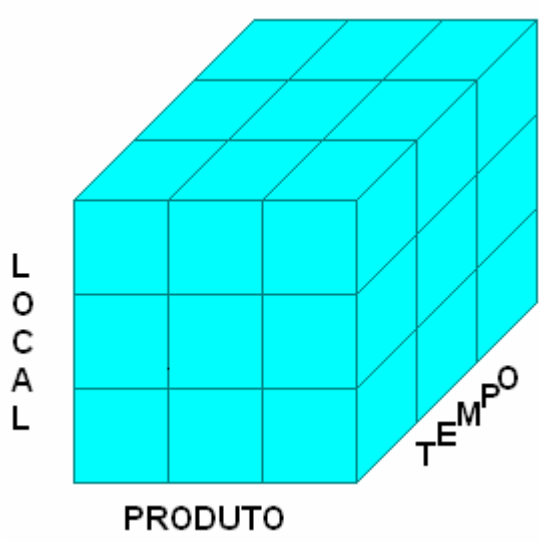


Figura 5 – Representação do modelo dimensional como um cubo de dados

Neste cenário são criadas tabelas de dimensões que terão suas chaves primárias ligadas ao fato vendas como chaves estrangeiras formando o modelo dimensional do *DW*, que também pode ser chamado de modelo estrela pois as dimensões estarão em torno do fato como mostra a Figura 6.

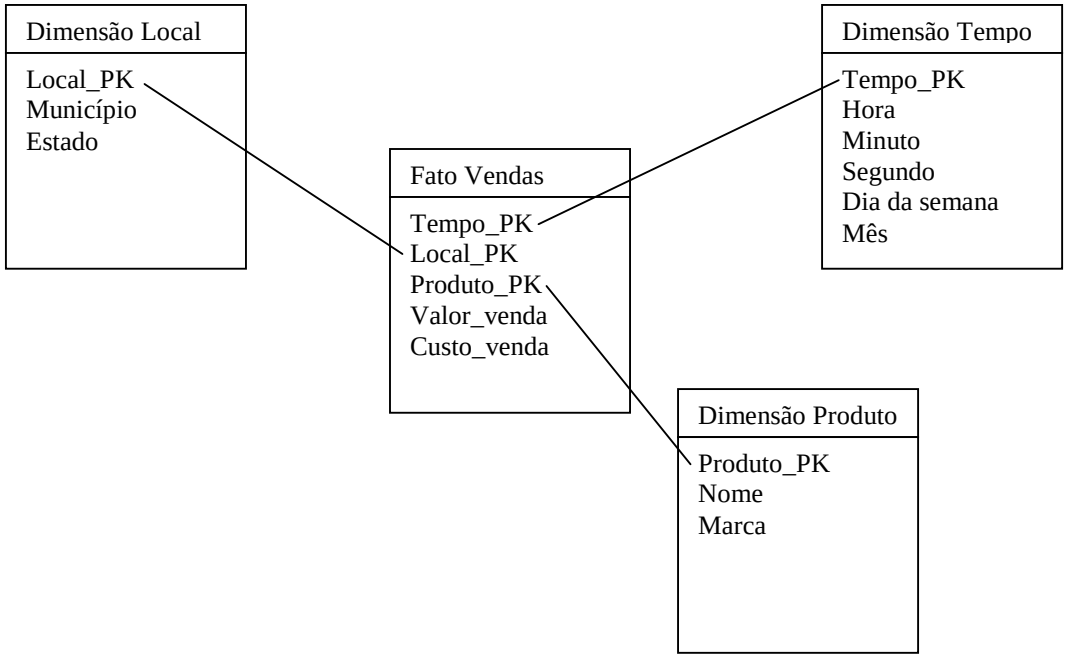


Figura 6 – Modelo dimensional do *DW* (estrela)

2.4.2 NORMALIZADO

Em sistemas OLTP o modelo normalizado é o mais utilizado porque deve armazenar as informações da organização com mais detalhes quanto for possível. Criado por Edgar F. Codd, nos anos 70, o Modelo Relacional, começou a ser utilizado nas empresas a partir de 1977. A abordagem relacional está baseada no princípio de que as informações em uma base de dados podem ser consideradas como relações matemáticas e que estão representadas de maneira uniforme, através do uso de tabelas bidimensionais.

Segundo Demarest [2001], a normalização tem muitas vantagens para manter a integridade dos dados de forma eficiente para o processamento on-line de transações, mas atua fortemente contra a legibilidade dos dados no ambiente de suporte à decisão, como segue:

- Gera um grande número de tabelas: o propósito da normalização é separar dados independentes em entidades distintas. É comum em atividades complexas, encontrar modelos com mais de mil tabelas.

- Há múltiplas formas de ligar duas tabelas, e a escolha do caminho faz grande diferença no resultado obtido e no tempo de resposta de uma consulta.

Em um ambiente de *DW* a normalização do modelo irá interferir diretamente na eficiência de todo o sistema. Uma tabela de fato já é totalmente normalizada mas a normalização das dimensões apenas irá aumentar o número de relacionamentos necessários para se obter uma resposta desejada de uma consulta requerida pelos usuários, prejudicando assim o tempo dessa resposta devido ao aumento de processamento para executá-la. Evidentemente diminui-se o espaço utilizados pelos dados mas em contrapartida é necessário a criação de mais estruturas de indexação para auxiliar as consultas, o que acaba gerando perda de espaço físico.

2.4.3 SNOWFLAKE

O modelo de dados "*Snowflake*" é uma derivação do modelo dimensional e normalmente é utilizado nos casos onde uma dimensão necessita que um de seus atributos seja melhor detalhado. Nesta situação é criada uma mini-dimensão que será ligada a dimensão através de uma chave artificial e que por sua vez manterá os dados detalhados do atributo (Ex. atributos demográficos em dimensão grandes).

Mas com raras exceções o uso desse modelo não deve ser recomendado em uma arquitetura de *DW*, pois a grande quantidade de relacionamentos entre as dimensões fará com que as consultas se tornem pouco eficientes, o que não compensa mesmo levando em consideração que neste tipo de modelo tende a diminuir o espaço físico de armazenamento, além disso este modelo acaba por se tornar muito detalhista o que intimida os usuários, como mostra a Figura 7.

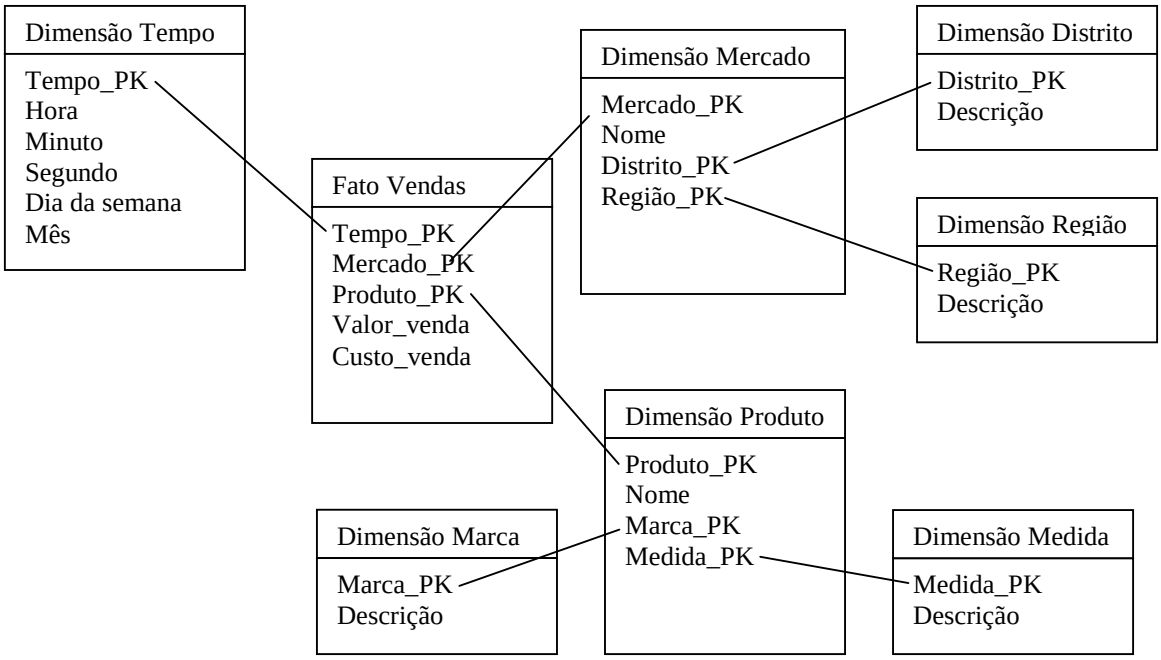


Figura 7 – Modelo *Snowflake*

2.5 *BACK-END*

O Projeto de *Back-End* engloba a criação do modelo físico de dados (Projeto Físico), todas as ferramentas e processos utilizados na obtenção dos dados dos sistemas operacionais ou transacionais e dados de origem externa, ou seja, a extração dos dados nos sistemas fontes, com as transformações adequadas, no tempo adequado e o armazenando na área de estágio (Processo de *ETL*), para posterior carga no *DW* onde serão indexados e agregados de maneira a refletir na performance das posteriores consulta (Estratégias de Otimização de Repositório), assim como a atualização do metadados que mantém a documentação desses dados desde a extração, as regras de limpeza e transformações, até o armazenamento, que são de suma importância para interação do usuário com o *DW*. Este processo é mostrado na Figura 8, e de acordo com especialistas, este processo demanda cerca de 80% do esforço de um projeto de implementação de um ambiente de data warehouse e, em muitos casos, ocupa muito mais tempo que o previsto, Campos & Filho [1997].

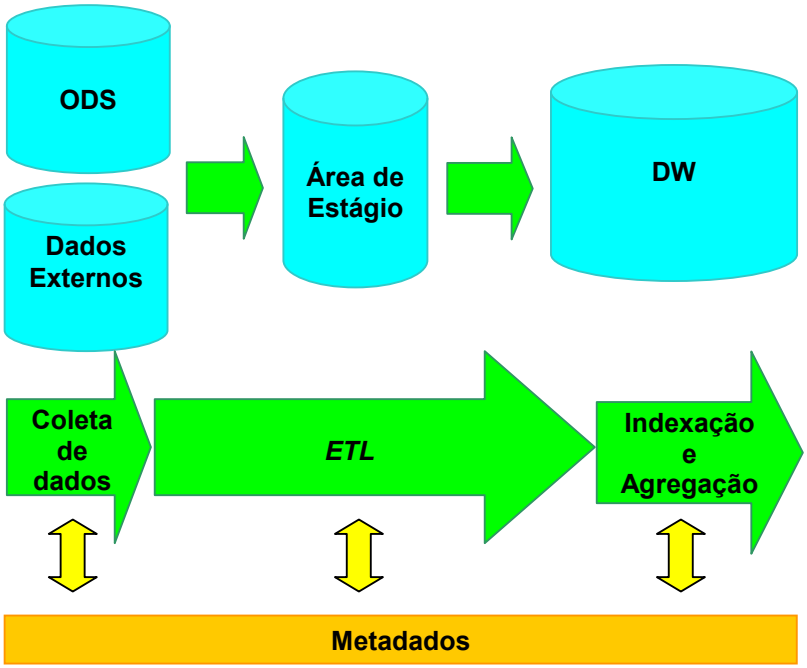


Figura 8 – Processo de *Back-End*

2.5.1 PROJETO FÍSICO

É a etapa inicial do projeto de *Back-End*, onde se define o sistema gerenciador de bancos de dados (SGBD) que melhor se adequar as necessidades para a construção do *DW*. Segundo Inmon [1997], o SGBD deve ter algumas características tecnológicas para o processamento satisfatório do *DW*. Estes incluem uma interface de linguagem robusta, suporte a chaves compostas e dinâmicas no seu comprimento de dados, e as habilidades de fazer o seguinte:

- Administrar quantias grandes de dados;
- Administrar dados em diversas mídias;
- Fácil indexação e monitoração de monitor;
- Interface com o maior número de tecnologias;
- Permita ao programador colocar os dados diretamente no dispositivo físico;

- Armazena e acessa dados com acesso paralelo;
- Tenha controle do metadado do *DW*;
- Carga eficiente no *DW*;
- Utilização eficiente de índices;
- Compactação de dados;
- Suporte a chaves compostas;
- Gerenciamento de Lock;
- Processamento de índices;
- Recuperação rápida de grandes volumes de dados.

Em seguida inicia-se a definição da padronização a ser utilizada no modelo físico que deve ser feita em conjunto com os usuários e amplamente divulgada. Kimball [1998], diz que para o data warehousing existem dois conjuntos importantes de padrões de nomenclatura: aqueles para nomes de objeto de banco de dados e aqueles para nomes de arquivo e locais físicos. São padronizados as nomenclaturas dos objetos de dados que os identifique de maneira a deixá-los de fácil compreensão, os tipos de dados que cada objeto irá armazenar, soluções para similaridades de nomes em diversos ambiente, a utilização de abreviações devido a possibilidade de limitação de tamanho pelo SGBD e por fim o nomes dos arquivos físicos e seus locais.

Definido os padrões de nomenclatura é necessário criar as chaves artificiais que substituirão as chaves do operacional, que podem aparecer como atributos de dimensão em modelos dimensionais mas não deviam servir como a chave primária das tabelas de dimensão, que sempre devem ser uma chave substituta, Kimball [1998], isto porque o ambiente operacional tende a sofrer menor periodicidade de armazenamento o que pode provocar a reutilização de chaves existentes gerando duplicidade e inconsistências no resultado das consultas. As chaves primárias e estrangeiras do modelo físico são importantes pois o *DW* é um ambiente de extrema utilização de consultas e quanto

melhor criadas as chaves melhor será a eficiência da resposta das consultas devido as várias junções entre as tabelas envolvidas.

O próximo passo é a estimativa do volume de dados inicial e do crescimento do *DW*, o segundo impacta diretamente na infra-estrutura de *hardware* escolhida. Em quase todos os armazéns de dados, o tamanho das tabelas de dimensões é insignificante comparado ao tamanho das tabelas de fato. A segunda maior coisa no ambiente é o tamanho dos índices das tabelas de fato, Kimball [998]. O cálculo do tamanho médio de cada registro, do tamanho dos índices iniciais a serem criados, o tamanho que ocupará a área de estagiamento, o espaço utilizado pelos agregados e a projeção de crescimento para cada item descrito, podem fornecer a estimativa de volume, mas é importante considerar uma margem de erro grande no inicio para que não ocorra imprevistos posteriormente.

2.5.2 PROCESSO DE *ETL*

Visto como um dos processos fundamentais para o sucesso de um *DW,* a extração dos dados de suas fontes origens, o tratamento realizados sobre esses dados e a carga dos mesmos, consome mais da metade de todo o processo de concepção de um *DW*. A maioria do esforço exigido no desenvolvimento de um *DW* é consumido neste momento e não é incomum que oitenta por cento de todo esforço seja empregado no processo de ETL, Inmon [1997]. Este processo pode exigir uma infra-estrutura suficientemente robusta para suportar tais etapas gerenciadas pelo metadados, pois esta terá que estar pronta para receber e processar possíveis grandes quantidades de dados obtidos nas origens podendo utilizar grandes áreas de estagiamento. É de suma importância que os dados encontrados no *DW* sejam de ótima qualidade, ou seja, íntegros, consistentes, não redundantes, confiáveis, e o mais atualizados e apresentáveis ao usuário do *DW,* determinando assim o sucesso de todo o projeto.

A extração dos dados leva mais ou menos 60 por cento das horas de desenvolvimento de um *DW* apenas no processo de extração, Kimball [1998] e deve se basear na busca das informações mais importantes em sistemas fontes ou externos e que estejam em conformidade com a modelagem do *DW*. Tal busca de dados pode ser

obstruída por problemas como a distribuição das origens dos dados, que podem estar em bases distantes com plataformas diferentes gerando a demanda de utilização de formas de extração diferentes para cada local. Dados que tendem a ser manipulados por diversos sistemas podem apresentar diferença de sua origem, deve-se extrair esses dados para o *DW* diretamente dos sistemas que os originou e em caso da existência de diversas versões em um mesmo sistema, deve-se usar aqueles que são mais atuais e que já tenham passado por processos de encerramento diário ou mensal.

No momento de criação do *DW* é comum uma carga de dados inicial que faça com que a extração busque todos os dados dos sistemas fontes, mas com o decorrer do tempo a extração deve estar preparada apenas para fazer cargas incrementais. É muito mais eficiente a carga incremental que carrega apenas os registros que foram alterados ou inseridos desde a carga inicial, Kimball [1998], ou seja, procure buscar apenas as informações que sofrem modificações desde a última carga, gerando assim menor tempo de processamento na extração e menor tráfego de dados na rede.

A etapa de transformação consiste na aplicação das regras de limpeza e transformação dos dados contidas no metadados. É proveniente deste momento a qualidade dos dados que irão popular o *DW,* e segundo Kimball [1998], os dados devem representar a verdade, a mais pura verdade, nada mais que a verdade. Para garantir a qualidade dos dados Kimball [1998] levanta como características mais relevantes:

a) Unicidade dos dados, evitando assim duplicações de informação;

b) Precisão dos dados, os dados não podem perder suas características originais assim que são carregados para o *DW*;

c) Dados completos, não gerando dados parciais de todo o conjunto relevante as análises; e

d) Consistência, ou seja os fatos devem apresentar consistência com as dimensões que o compõem.

Grandes problemas são encontrados no decorrer deste processo, como a ambigüidade de dados que podendo ter a mesma nomenclatura tem significados

diferentes, ou ao contrário, valores diferentes apresentarem o mesmo significa como os valores nulos que podem ser representados com conteúdo vazio ou 0 (zero). Existem problemas com a integridade referencial dos dados, onde dados que farão a composição de um fato podem ser negligenciados ou não encontrados na origem no momento de gerar uma dimensão. A representação de conjunto de caracteres em bases distribuídas e que são configuradas com tabelas de caracteres diferentes também implicam em problemas no momento da transformação. Por fim temos a aplicação das regras de cálculos existente nos metadados para gerar novos valores a partir de valores da origem, como por exemplo sumarizações de vendas.

A última etapa do processo de *ETL* é a carga em si dos dados já extraídos e transformados para dentro do *DW*, que podem estar armazenados em uma área de estagiamento. Basicamente serão carregadas as dimensões estáticas, de modificação lenta ou remanescentes e fatos integrantes ao modelo do *DW*. Este processo pode ter alto custo de processamento e que implica em tempo de carga que na maioria das vezes não pode ser extenso devido a utilização contínua do *DW*, assim algumas precauções devem ser tomadas antes de se iniciar a carga dos dados, como:

a) desligamento de índices e referencias de integridade (isso pode prejudicar na qualidade dos dados pois apesar de diminuir o processamento os dados não são validados no momento da inserção);

b) utilização de comandos do tipo truncate ao invés de delete pois nos SGBDs mais atuais este recurso não gera armazenamento de informações em áreas de recuperação de dados;

c) ter em mente que no momento da carga alguns dados não serão carregados e deste modo os mecanismos de carga devem dar suporte a auditorias de carga para que a carga possa ser re-iniciada no momento em que foi parada e a possibilidade de manter logs com os dados rejeitados para a avaliação dos motivos pelo qual não foram carregados e assim ajustados para integrarem o conjunto a ser carregado.

Dimensões estáticas normalmente não oferecem problemas, pois estas mantém dados que não sofrem alteração na sua origem e serão carregados uma única vez, assim

como as remanescentes que normalmente são originadas de esforço manual na sua confecção, por exemplo as planilhas eletrônicas. Já as dimensões de modificação lenta necessitam da verificação em suas fontes e nas auditorias das cargas para que se possa identificar qual o momento seguinte depois da última carga que deve iniciar o processo, gerando processamento na leitura de logs de sistemas operacionais e comparação de atributos, podendo então ser necessário sobrescrever todo o conteúdo de um registro, gerar um novo registro na dimensão ou criar um atributo a mais para armazenar o valor antigo, Kimball [1998].

As dimensões estando corretamente carregadas, pode-se iniciar a carga dos fatos que depois de modelados para a conter apenas os dados de importância para a organização direcionam quais regras serão utilizadas como por exemplo filtros do que será inserido ou somas a serem realizadas, provocando o aparecimento de regras que passaram despercebidas no início da modelagem.

Fatos demandam cuidado na sua carga como o uso das chaves artificiais das dimensões para que se tenha uma integridade referencial, controle de valores nulos obtidos no momento da transação para que não gerem a falta de integridade referencial como datas que estando nulas invalidarão o histórico do fato. Técnicas para amenizar o processo devido ao grande volume de dados podem ser usadas, como a carga incremental dos fatos, que irá carregar apenas dados novos ou alterados, execução do processo em paralelo e em momentos de pouco ou nenhum uso do SGBD e a utilização de tabelas auxiliares que serão renomeadas como definitivas ao fim da carga, Kimball [1998].

Com a finalização do processo de carga, é necessário a divulgação para os usuários da existência de dados atualizados no *DW*.

2.5.3 ESTRATÉGIAS DE OTIMIZAÇÃO DE REPOSITÓRIO

A etapa de otimização constitui na eficiência que o *DW* deve ter, ou seja responder com rapidez as consultas requeridas pelos usuários, e para obter este resultado usas-se os recursos oferecidos pelos SGBDs atuais, como indexação utilizando Árvore-B, Bitmaps, Hash ou formas de indexação proprietárias que fazem complexas consultas retornarem seu

resultado em tempo muito baixo ou suficientemente aceitável pelos usuários. Normalmente as chaves primárias e estrangeiras são indexadas com estruturas Árvore-B devido ao seu padrão de cardinalidade, mas em casos de atributos de baixa cardinalidade é sugerido o uso de indexação baseada na estrutura Bitmap.

Outra forma de indexação que vem demonstrando ser altamente eficiente são as bibliotecas para incorporar recuperação de informação, que mantém uma estrutura proprietária externa ao SGBD de forma a manter apenas uma ligação pelas chaves primárias das dimensões ou fatos do *DW*, assim o processamento fica todo em varrer a biblioteca e quando encontrados os registros válidos para a consulta basta apenas buscá-los diretamente dentro da estrutura do *DW* através de suas chaves primárias. Levando em consideração que tabelas de fatos tendem a ter índices apenas pela chave constituída das chaves das dimensões, as tabelas de dimensão podem ser indexadas por muitos, senão todos os atributos, Kimball [1998]. Mas a criação de muitos índices leva ao consumo de espaço físico de armazenamento, que muitas vezes pode ocupar mais do que os próprios dados.

Outra maneira de otimizar o repositório do *DW* é a utilização de tabelas agregadas que oferecem ganhos significativos de performance pois tendem a ser 100, 1000 ou mais vezes menores que as tabelas de fatos, Kimball [1998]. Assim um fato de vendas com granularidade de vendas por hora, pode gerar tabelas agregadas sumarizadas por vendas mensais ou anuais evitando assim esse tipo de informação requerida ser processada diretamente no fato de vendas mas sim nas tabelas de agregados que manterão menor número de registros oferecendo menor tempo de respostas para as consultas.

Para gerar as tabelas agregadas deve-se definir o que será agregado baseando-se nos requisitos das consultas mais freqüentes e a distribuição estatística dos dados, Kimball [1998]. Em suma os agregados devem oferecer performance ao usuário, minimizar o custo de manutenção na extração e carga dos dados e não sobrecarregar o administrador de base de dados com a manutenção dos agregados utilizando técnicas proprietárias do SGBD para construir agregados de forma a deixá-los transparentes as aplicações.

2.6 *FRONT-END*

O *Front-End* é o resultado visual do *DW*, são as informações que o usuário de negócio acessa e utiliza no seu cotidiano profissional, Kimball [1998]. Basicamente são as ferramentas que acessam os dados do *DW* e as disponibilizam visualmente para a interatividade com o usuário e fornecem os tratamentos de análises, estas ferramentas são conhecidas como ferramentas de processamento analítico *on-line* ou *OLAP* (*on-line analytical processing*). Tais ferramentas podem ser *ROLAP* (*OLAP* Relacional) que é um conjunto de interfaces para o usuário e aplicações que dão aos SGBDs relacionais uma aparência de dimensional, Kimball [1998], podem ser caracterizadas como *MOLAP* (*OLAP* multidimensional), que são um conjunto de interfaces, aplicações e tecnologias de SGBDs proprietários que tem aparência multidimensional, Kimball [1998], e por fim as *HOLAP* (*OLAP* híbrido) que combinas as ferramentas *ROLAP* e *MOLAP*.

Deve ser importante na escolha de ferramentas ou desenvolvimento das próprias ferramentas, que elas tenham a habilidade de gerenciar em qual estrutura de dados buscar as informações quando requeridas pelo usuário, ou seja, deve ser transparente através da ferramenta ou do SGBD se uma consulta vai buscar os dados em uma tabela agregada ou em um tabela de fato, escolhendo a que irá responder mais rapidamente. É importante também ter disponibilizado ferramentas de acesso e segurança aos dados, assim como de monitoramento das atividades do *DW*, para se analisar o uso e performance e mostrar resultados para a gerência. E no atendimento dos usuários de relatórios, é necessário que as ferramentas possam ser gerenciáveis a ponto gerar relatórios flexíveis, com possibilidade de passagens de vários parâmetros ou programação para execução, utilização de tabelas e gráficos, e que atendam a maioria das consultas requeridas. Isso pode gerar um grande processamento, e para amenizar, o ideal é a utilização de um servidor de relatórios.

E por fim essas ferramentas devem permitir operações de "*drill-down*" onde se pode chegar ao nível de detalhamento mais alto dos dados e operações de "*roll-up*" que significa sumarizar os dados no nível mais baixo de detalhamento, além de disponibilizar o acesso web para facilitar a sua utilização.

2.7 CONCLUSÃO DO CAPÍTULO

Este capítulo abordou aspectos da construção de um *DW*, desde a modelagem da sua estrutura, criação física e disponibilização do seu conteúdo ao usuário final, onde fica claro a sua importância para se obter informações pertinentes ao processo decisório das organizações, e também sendo o *DW* um diferencial de competitividade altamente relevante para o sucesso das empresas.

3 O *DATA WAREHOUSE* DO CNPQ

3.1 INTRODUÇÃO

Criado pela Lei nº 1.310 de 15 de janeiro de 1951, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) é uma Fundação, vinculada ao Ministério da Ciência e Tecnologia (MCT), para o apoio à pesquisa brasileira. Contribuindo diretamente para a formação de pesquisadores (mestres, doutores e especialistas em várias áreas de conhecimento), o CNPq é, desde sua criação, até hoje, uma das maiores e mais sólidas estruturas públicas de apoio à Ciência, Tecnologia e Inovação (CT&I) dos países em desenvolvimento, CNPq [2005].

Os investimentos feitos pelo CNPq são direcionados para a formação e absorção de recursos humanos e financiamento de projetos de pesquisa que contribuem para o aumento da produção de conhecimento e geração de novas oportunidades de crescimento para o país, CNPq [2005].

Com a intenção de tornar compatível e integradas as informações em toda a interação entre o CNPq com seus usuários, foi criado um conjunto de sistemas de informação denominados “Plataforma Lattes” que tem como objetivo melhorar a qualidade dessas informações e sua estrutura para o melhor entendimento dos pesquisadores e estudantes que a disponibilizam.

A Plataforma Lattes representa a experiência do CNPq no que se refere à integração de seus sistemas de informações gerenciais, instrumento fundamental não só para as atividades de fomento operadas pela Agência mas também para tratamento e difusão das informações necessárias à formulação e à gestão de políticas de ciência e tecnologia, LATTES [2005].

Nenhuma instituição é capaz de cumprir de maneira eficaz a sua missão sem um conhecimento atualizado da realidade de seu objeto ou campo de trabalho. É justamente com esse propósito que o CNPq vem realizando um enorme esforço no sentido de integrar as suas bases de informações, LATTES [2005].

Essa integração tem como fonte primária de coleta de dados quatro projetos distintos, porém integrados, LATTES [2005].

O primeiro deles se refere a um Sistema Eletrônico de Currículos. O registro da vida pregressa e atual dos pesquisadores é elemento fundamental para a análise de seu mérito e competência. Nesse domínio, o Brasil logrou desenvolver um formato-padrão para coleta de informações curriculares, adotado não só pelo CNPq mas pela maioria das agências de fomento do País, LATTES [2005].

A adoção de um padrão nacional de currículos, com a riqueza de informações que esse sistema possui, a sua utilização compulsória a cada solicitação de financiamento e a sua disponibilização pública, na internet, deram muito mais transparência e confiabilidade às atividades de fomento do CNPq. Nesta base, hoje, constam mais de 200 mil currículos atualizados, LATTES [2005].

O segundo sistema é o Diretório dos Grupos de Pesquisa no Brasil. O Diretório é uma base de dados que registra todos os grupos de pesquisa em atividade no País. As informações constantes na base de dados dizem respeito aos recursos humanos engajados no grupo, às linhas de pesquisa em andamento, às especialidades do conhecimento, aos setores de aplicação, aos cursos de mestrado e doutorado com os quais os grupos interagem e à produção científica e tecnológica captada a partir do sistema eletrônico de currículos, LATTES [2005].

O terceiro sistema é o Diretório de Instituições. O registro acurado das instituições que demandam fomento ao CNPq, ou que tenham membros participantes dos grupos de pesquisa, ou que ofereçam cursos de graduação ou pós-graduação, é fundamental para que não somente o CNPq mas também as agências de fomento

brasileiras e os órgãos encarregados do planejamento e acompanhamento do desenvolvimento da Ciência e Tecnologia no Brasil possam ter um mapa preciso da distribuição de recursos e da localização da competência de pesquisa e desenvolvimento no País e no exterior, LATTES [2005].

O quarto sistema chama-se Sistema Gerencial de Fomento. Este sistema é imprescindível para uma gestão estratégica e para dar mais qualidade às atividades de fomento do CNPq, LATTES [2005].

Esta plataforma é baseada na metodologia de desenvolvimentos de projetos E-Gov proposta por PACHECO [2003], que tem como formação de sua base as unidades de informações, subindo para a camada de padronização, sistematização e publicação de informações e serviços, até chegar ao mais alto nível que mantém a gestão, produção e publicação de conhecimento, mostrado na Figura 9.

Figura 9 – Arquitetura conceitual para projetos de E-Gov.

Unidades de Informação: as unidades de informação descrevem subdomínios da área-fim para a qual a plataforma está sendo desenvolvida. São formadas por classes ou elementos do domínio da plataforma para os quais estão associados conteúdo, processos e serviços específicos. Uma unidade de informação não pode ser genérica a ponto de não especificar conteúdo e processos independentes nem ser específica na

descrição ou funcionamento (casos em que provavelmente seja um elemento da unidade de informação).

Os principais elementos metodológicos e tecnológicos da camada de unidades de informação estão relacionados à padronização do conteúdo de cada unidade. Aos responsáveis pelo projeto caberá identificar as unidades, especificar conteúdo inicial (de acordo com as diretrizes da fase de Projeto), propor um padrão inicial para essas unidades e promover sua contínua revisão por parte da comunidade interessada. Para proceder ao estabelecimento desses padrões, é necessário definir a ontologia[1] das unidades de informação e sua padronização XML.

Padronização XML: a tarefa de estabelecer ontologias para as unidades de informação da plataforma deve produzir como resultado padrões compartilháveis e intercambiáveis entre os interessados (preferencialmente partícipes de uma comunidade virtual de padronização). Para tal, é fundamental que o projeto da plataforma inclua a produção explícita desses padrões. Esta se constitui na formação de metadados para o domínio da plataforma, os quais são especificações criadas por pessoas (ou geradas de forma automática por computadores), descritas na forma de estruturas e regras de manejo da informação, e que permitem que pessoas e computadores interajam com as informações especificadas pelos metadados.

Fontes e Sistemas de Informação: a camada de fontes e sistemas de informação inclui os repositórios de cada unidade de informação da plataforma e os respectivos sistemas de informação que captam, tratam e armazenam os dados da unidade junto à comunidade usuária. O desenvolvimento desses componentes tecnológicos tem por base as padronizações da camada inferior e seguem metodologia e as tecnologias estabelecidas na fase de projeto da plataforma.

Portais e Serviços Web: a terceira camada da plataforma é composta pelos instrumentos desenvolvidos para apresentação de informações na Web (websites), para

[1] Ontologias estabelecem compreensão compartilhada e comum de um domínio e podem ser trocadas entre pessoas e computadores (Studer et al., 2000).

publicação de informações dinamicamente atualizadas com interação com a comunidade usuária (portais Web) e pelos recursos de disseminação de serviços de informação de governo na Web (Web services).

Sistemas de Conhecimento: finalmente, no topo da arquitetura da plataforma de governo estão os sistemas de conhecimento. Trata-se dos instrumentos projetados para gerar novos conhecimentos a partir das fontes de informação da plataforma e de sua operação por parte da comunidade usuária. Para elaborar os instrumentos dessa camada, os responsáveis pelo projeto deverão valer-se da área de descoberta de conhecimento[1] e, a partir de técnicas relacionadas (e.g., mineração de dados, estatística, reconhecimento de padrões), elaborar projetos específicos voltados à produção de conhecimento sobre área de governo atendida pela plataforma.

Para atender a infra-estrutura de informação necessária a análise dos dados de C&T nacionais de forma integrada e uniforme levando em consideração a necessidade dos diferentes usuários dessa informação, foi criado o *Data Warehouse* da Plataforma Lattes, composta pelos *Data Marts* de Fomento, Grupos de Pesquisa e de Currículo, e que através do estudo de seus modelos de dados e seus processos de *Back-End* e *Front-End*, pode-se observar a ocorrência dos indicadores propostos para a avaliação técnica de projetos de *data warehouse.*

3.2 A ARQUITETURA DO *DATA WAREHOUSE* DO CNPq

O *Data Warehouse* do CNPq é composto por *data marts* integrados que foram concebidos para agregar os dados associados ao fomento (bolsas, auxílios integrados e passagens), aos currículos dos pesquisadores e estudantes e ao diretório de grupos de pesquisa. Com o cruzamento dos dados gerados por esses *DM* se obtém uma visão das atividades de pesquisa e a resposta da aplicação do fomento na produtividade e formação dos pesquisadores.

---

[1] Termo cunhado para salientar o produto final de um banco de dados deve ser a descoberta de conhecimento (Fayyad et al. , 1996).

Através de um processo de *ETL* os dados oriundos do fomento, currículos e diretório de grupos de pesquisa são depositados em seus respectivos *DMs* para então serem disponibilizados através de componentes de apresentação como representa a Figura 10.

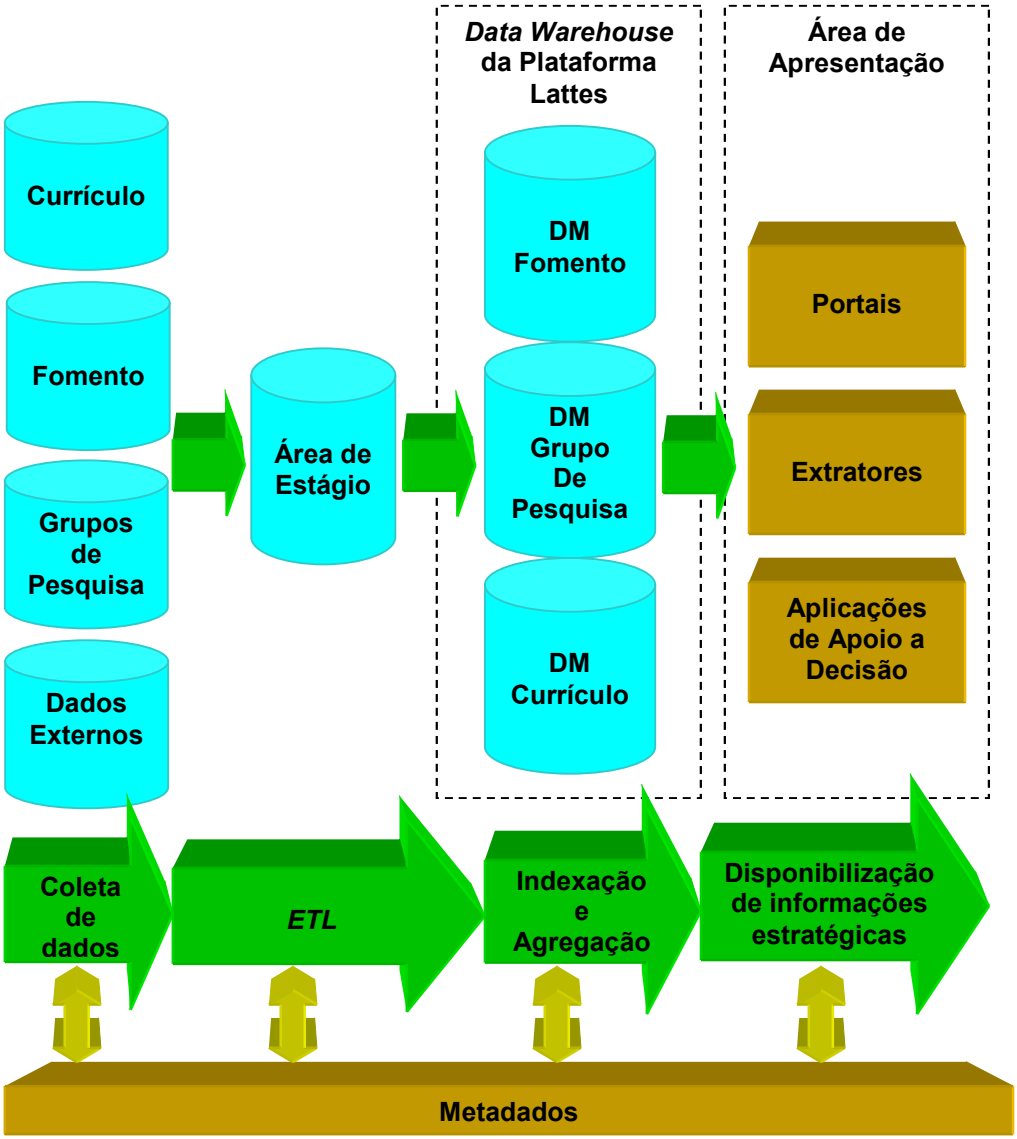


Figura 10 – Processo de carga e publicação dos dados

3.3 O *DATA MART* DE FOMENTO

Como principal atividade do CNPq o fomento gera a criação de programas de apoio a formação de recursos humanos e à pesquisa e desenvolvimento de atividades

relacionadas a ciência e tecnologia, CNPq [2005]. Com o objetivo de buscar essa informação e organizá-la em um repositório para atender as necessidades de análise de seus usuários, criou-se o *DM* de Fomento, que disponibiliza análises de alocação, fluxo e adequabilidade de lotação de recursos nas ações de fomento do CNPq, tudo isso obtido através de informações vindas de sistemas que mantém os currículos, grupos, projetos, produções, instituições e outros.

3.3.1 MODELAGEM

Para se obter as análises quantitativas o modelo de dados do *DM* de Fomento contém dimensões e fatos de extrema importância que podem ser consultados no Anexo 1, mas dentre eles podemos citar algumas a seguir na tabela 1.

Tabela 1 – Dimensões e Fatos do *DM* de Fomento

Dimensões e Fatos	Descrição	Alguns campos
Pessoas	mantém os dados dos pesquisadores e estudantes	nome, sexo, endereço, e-mail
Área do conhecimento	área científica reconhecida pelo CNPq, pode ser a área de atuação de alguém que recebe bolsa, ou a área de formação de um pesquisador	código da área, descrição da área
Modalidade	tipo de bolsa que um recurso humano recebe	Nome, classificação, nível da categoria
Instituição	são as organizações onde se formaram, ou que fazem a gestão dos recurso disponibilizados pelo CNPq	nome, se é agência de fomento, se participa do diretório de pesquisa

Fundos Setoriais	investimentos realizados pelo fundo setorial	nome, descrição
Pagamento	fato que mantém todos os pagamentos realizados, bolsa, auxílio, despesas de viagem	chaves das dimensões Pessoa, Instituição, Modalidade e Área de Conhecimento, e Quantidade de itens de despesa, valor total dos itens

Estas tabelas e outras mais formam o modelo dimensional do tipo estrela, onde o fato Pagamento é o centro do modelo e em seu redor dimensões como Pessoas, Área do conhecimento, Modalidade e Instituição, como mostra a Figura 11.

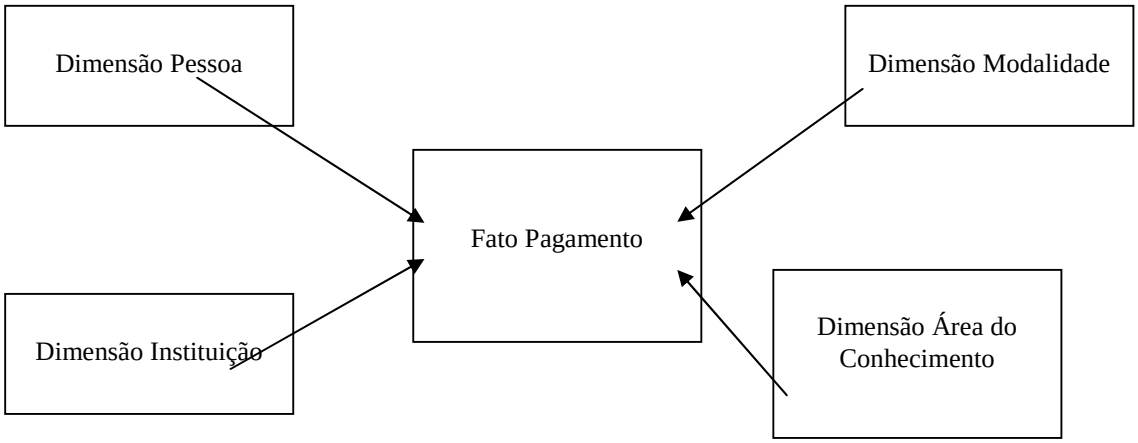


Figura 11 – Modelo estrela simplificado do *Data Mart* de Fomento

3.3.2 *BACK-END*

Toda a infra-estrutura de *back-end* do *DM* de Fomento está baseada em aplicações de cargas desenvolvidas em linguagem própria do sistema gerenciador de

banco Oracle®, onde foram desenvolvidos pacotes contendo procedimentos e funções que gerenciam toda a busca dos dados de origem a partir dos sistemas operativos do CNPq e do PROSSIGA. Esses fazem todo o processo de transformação e limpeza dos dados se utilizando de uma área de estagiamento, para então depois serem carregados em definitivo nos repositórios finais onde um mantém as tabelas de dimensões e outro as tabelas de fatos. Toda a carga é feita uma vez ao mês devido ao fechamento do pagamento de bolsas e auxílios, e esta se utiliza de tabelas de auditorias que indicam quando uma carga de dimensão ou fato iniciou, quando terminou e se foi bem sucedida, desta forma possibilitando o reinicio de uma carga que apresentou problemas de onde parou ou fornecendo informação sobre ocorrências que não devem interromper toda a operação durante o processo através de envio de e-mail aos responsáveis. Além da criação de índices para aumentar a performance nas consultas, foi usado o recurso de criação de visões materializadas do SGBD Oracle®, assim vários agregados foram criados em um repositório próprio com base nos dados dos fatos e o próprio SGBD decide se em uma determinada consulta irá buscar a informação no próprio fato ou se é mais performático o uso de um agregado nas visões materializadas.

3.3.3 *FRONT-END*

Foram desenvolvidas ferramentas de *Front-end* baseadas em programação JAVA® que buscam os dados fornecidos através do *DM* de Fomento e apresentam o investimento em Ciência, Tecnologia e Informação do país, realizados pelo CNPq. Estas ferramentas estão disponíveis para o público em geral no site http://fomentonacional.cnpq.br/dmfomento/home/index.jsp. As consultas estão divididas entre investimento em recursos humanos, investimento à pesquisa, mantendo-se também os investimentos do CNPq de forma consolidada nas linhas de atuação de bolsa no país e exterior, e fomento a pesquisa, assim como o detalhamento desses mesmos investimentos através dos fundos setoriais mostrado na Figura 12. Cada modalidade de consulta oferece diversos cortes na busca como: Bolsa no País por Área de conhecimento, Faixa etária, modalidade, etc, além de cruzamentos entre si como: Área de conhecimento por Faixa etária, ou por modalidade, etc, mostrado na Figura 13.

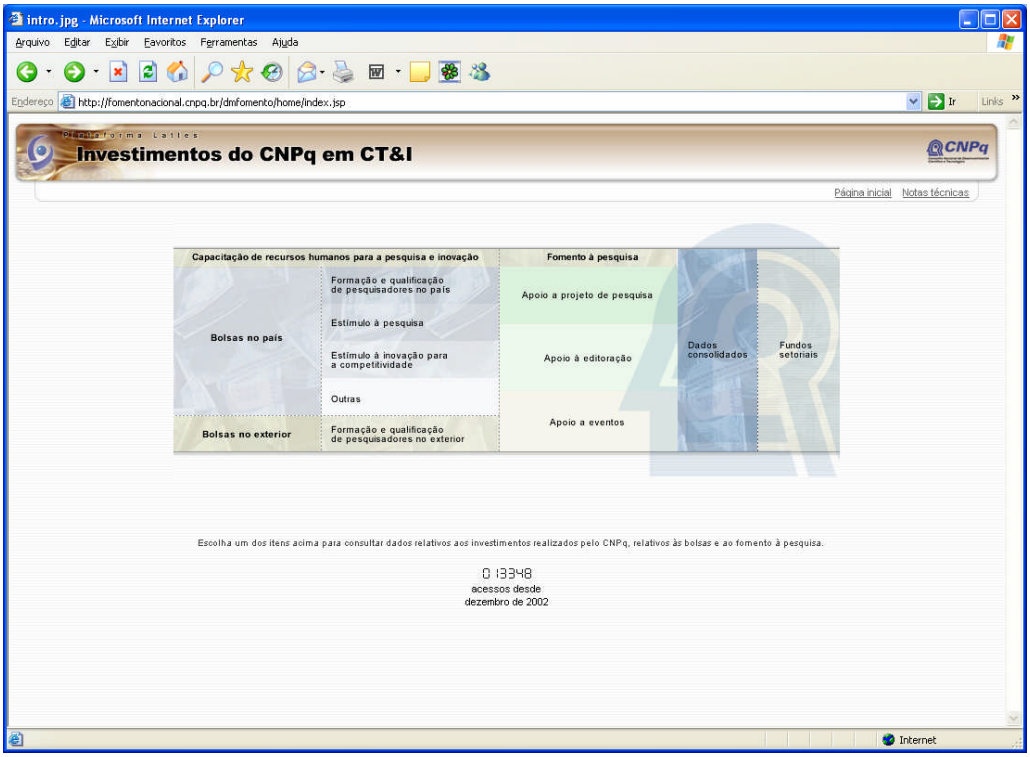


Figura 12– Site do *Data Mart* de Fomento

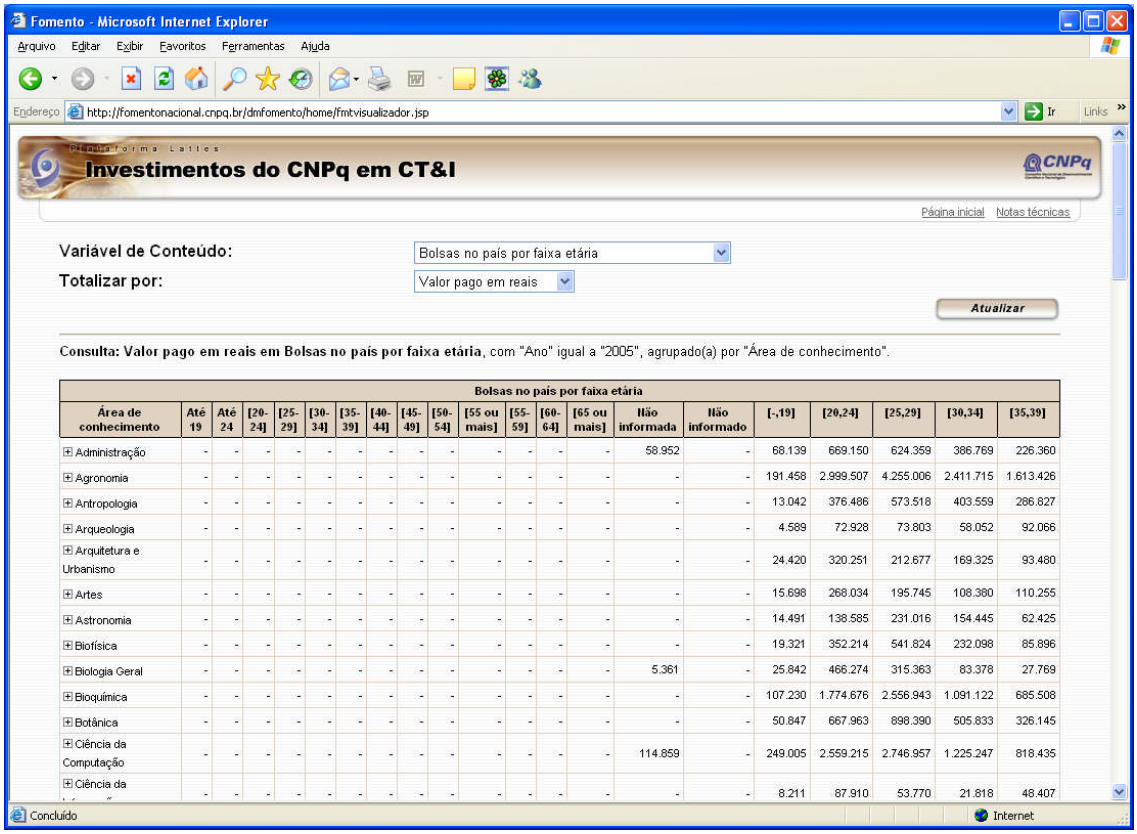


Figura 13 - Site do *Data Mart* de Fomento

3.4 O *DATA MART* DE GRUPOS DE PESQUISA

Para suprir a necessidade de análise de informações relativas a evolução dos censos de pesquisas e aos dados dos currículos dos integrantes dos grupos de pesquisa ao longo do tempo, foi projetado o *DM* de Grupos de Pesquisa. Com os dados desde repositório é possível ter a visão evolutiva censitária e temporal dos grupos de pesquisa, ou seja, pode-se avaliar as informações geradas no censos de pesquisa já realizados como distribuição percentual de grupos por região, e as informações relacionadas aos grupos ou participantes dos grupos, como análises sobre datas de criação de grupos ou publicações dos participantes. Este *DM* foi desenvolvido utilizando a proposta de Tissot [2004], que propõem um conjunto de artefatos para o direcionamento do levantamento de requisitos e a utilização de uma linguagem mais próxima daquela utilizada no cotidiano dos usuários.

3.4.1 MODELAGEM

O modelo de dados *DM* de Grupos de Pesquisa é composto por várias dimensões conformadas, já vistas no *DM* de Fomento, e novas dimensões e fatos que possibilitam as análises relacionadas à evolução temporal e censitária dos grupos, estas e as outras entidades relacionadas ao modelo podem ser vistas no Anexo II. Entre as novas dimensões e fatos podemos destacar as seguintes na tabela 2:

Tabela 2 - Dimensões e Fatos do *DM* de Grupos de Pesquisa

Dimensões e Fatos	Descrição	Alguns campos
Atividade Técnica	atividades técnicas vinculados as grupos de pesquisa	nome das atividades
Característica Grupo	mantém dados do período de criação e dissolução dos grupos	faixa temporal de criação do grupo, ano inicial, ano final
Palavra Chave Linha	está dimensão faz uma	chaves da dimensão

Pesquisa	"ponte" para a dimensão Palavra Chave, afim de relacionar as palavras chaves do grupo por linhas de pesquisas	Palavra Chave e fato de linha de pesquisa
Linha Pesquisa	fato que mantém todas as linhas de pesquisas que o grupo atua	Nome da linha de pesquisa, total de pesquisadores, estudantes, objetivo da linha de pesquisa
Técnico Grupo	este fato armazena todos os dados relevantes as atividades técnicas dos grupos de pesquisa	região geográfica, área de conhecimento, totais de mestres e doutores ligados a atividade técnica do grupo

O *DM* de Grupos de Pesquisa foi modelado como um modelo dimensional estrela, mas com a característica de se utilizar de dimensões "pontes", que auxiliam na ligação de fatos que mantém atributos que podem ser referenciados mais do que uma vez para um mesmo grupo de pesquisa em uma determinada dimensão. Por exemplo, a ligação entre o fato Linha Pesquisa com a dimensão Palavra Chave, onde para armazenar as diversas palavras chaves relacionadas a cada linha de pesquisa foi criada a dimensão "ponte" Palavra Chave Linha Pesquisa, como mostra a Figura 14.

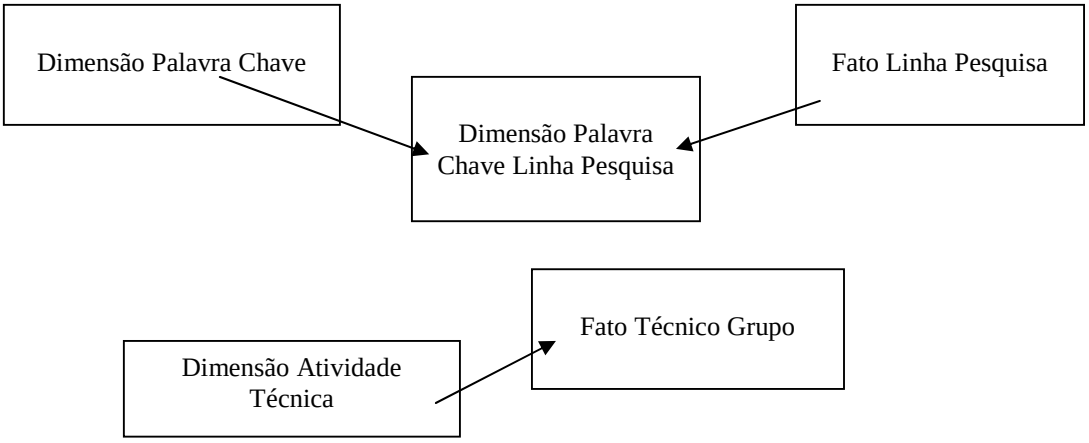


Figura 14 – Modelo parcial do *DM* de Grupo de Pesquisa

3.4.2 *BACK-END*

Todo o modelo de dados do *DM* de Grupos de Pesquisa é baseado no SGBD Oracle®, o processo de *ETL* foi desenvolvido utilizando linguagem de programação JAVA®. A origem dos dados vem dos sistemas legados do CNPq que já mantinham dados de censos anteriores que então são preparados e carregados no repositório específico para o *DM* de Grupos de Pesquisa. A carga dos dados desde repositório não necessita de área de estagiamente devido ao seu pouco volume, e problemas ocasionados durante o processo são armazenados em uma tabela de log que é acompanhada diariamente. Como o processo se realiza buscando atualizações feitas nos grupos, todo dia o processo é executado, fazendo a carga dos grupos novos, atualizados e possíveis grupos que apresentaram problema na última carga. Para auxiliar a performance das ferramentas *OLAP* desenvolvidas para acessar os dados do *DM* de Grupos de Pesquisa, foram analisadas as consultas necessárias para a busca da informação e então criados índices capazes de atender várias consultas que se utilizam de mesmos fatos e dimensões.

3.4.3 *FRONT-END*

Para disponibilizar os dados do *DM* de Grupo de Pesquisa, foram criadas ferramentas como o site http://dgp.cnpq.br/censo2004/ , que mantém os censos de 2000, 2002, 2004 e o atual, e torna possível análises quali-quantitativas a partir de séries históricas com informações que sintetizam a evolução temporal e agregada do perfil dos grupos de pesquisa oferecendo cortes por área de conhecimento, sexo e região geográfica – Figura 15, além de se obter súmulas estatísticas que fornecem um retrato bastante nítido da capacidade instalada de pesquisa no país, e ainda análises quantitativas do perfil da pesquisa através de cruzamentos de variáveis em um plano tabular, como mostra a Figura 16. Estas ferramentas utilizaram a metodologia de implementação de instrumentos de análise multidimensional da informação proposta por Gonzaga [2005].

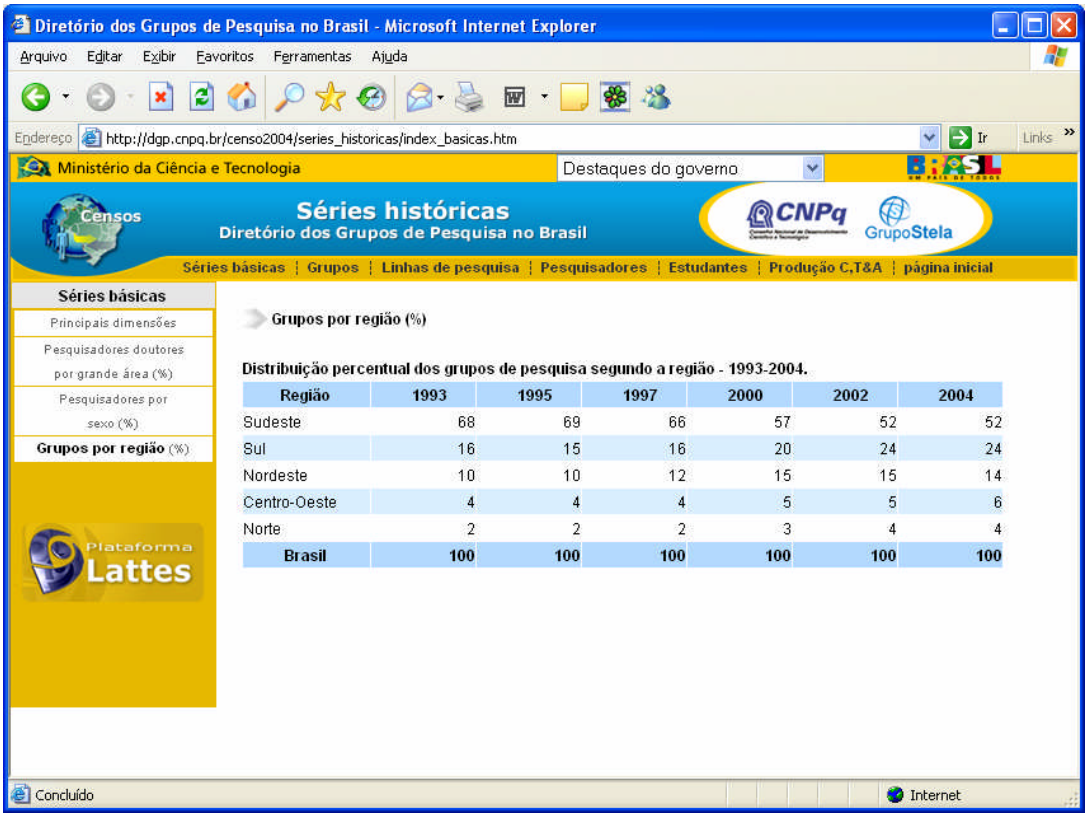


Figura 15 – Séries históricas de grupos por região

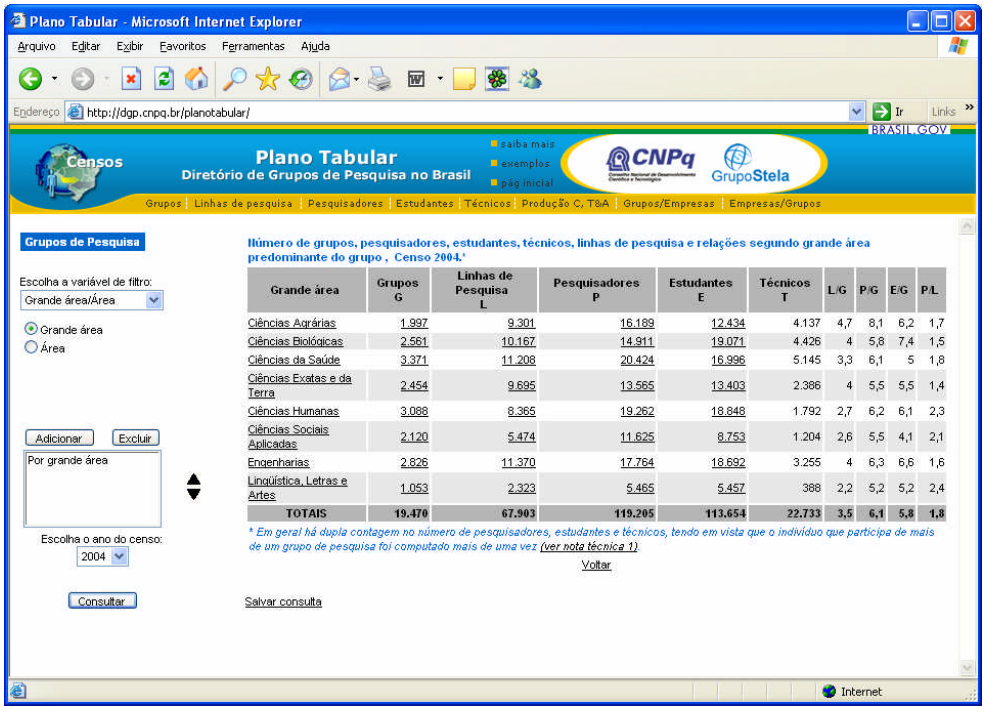


Figura 16 – Plano tabular de grupos de pesquisa por área de conhecimento

3.5 O *DATA MART* DE CURRÍCULOS

Como base integrante dos outros *DM*s, o *Data Mart* de Currículos visa fornecer subsídios para os instrumentos de análise de C&T, que são a investigação de conhecimento, estratificação de grupos de pesquisa e pesquisadores assim como o relacionamento entre as unidades de análise. Além da possibilidade de buscas textuais diretamente nos currículos armazenados no *DM* para se obter conhecimento sobre pesquisadores em suas áreas de atuação é possível também obter análises de demografia curricular apresentando o perfil da pesquisa quanto a distribuição da população pesquisada e análises de egressos de curso de graduação, especialização, mestrado, e doutorado no país entre outras análises.

3.5.1 MODELAGEM

O modelo de dados do *DM* de Currículo mantém as principais dimensões conformadas de todo o *DW*, e seu modelo dimensional tem como centro a dimensão pessoa. Todo o modelo do *DM* pode ser visto no Anexo III, mas em particular podemos destacar as dimensões e fotos da tabela 3.

Tabela 3 - Dimensões e Fatos do *DM* de Currículos

Dimensões e Fatos	Descrição	Alguns campos
Pessoa	contém os dados cadastrais dos pesquisadores	data de nascimento, sexo, endereço e e-mail
Produção científica e tecnológica	contém as informações relativas à produção científica	tipo e subtipo da produção, ano, país de publicação, idioma, título, meio de publicação, etc
Atividade profissional	contém as informações históricas da atividade profissional do pesquisador, sejam elas	instituição, ao período de início e término de cada atividade, tipo de vínculo e enquadramento funcional

	acadêmicas ou privadas	
Área de conhecimento	composta das informações referentes às áreas de conhecimento associadas a atuação, à formação, às linhas de pesquisa e à produção científica do pesquisador	Nome da área, nome da grande área, nome da especialidade, se está ativa

A chave primária da dimensão pessoa faz partes de vários fatos e também de várias dimensões, tornando-se assim o centro do modelo do *DM* de Currículo, como mostra a Figura 17, e junto com as outras unidades é possível a análise quantitativa de indicadores referentes a produção científica e tecnológica e artístico-cultural do país.

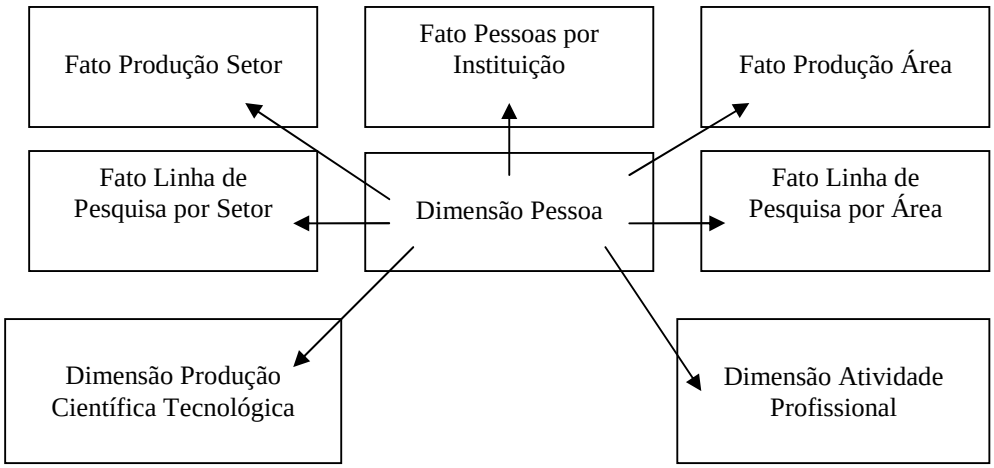


Figura 17 – Modelo parcial do DM de Currículo

Devido aos vários assuntos que foram contemplados pelo *DM* de Currículo, este modelo apresenta uma característica de desenvolvimento baseada na construção de vários agregados, ou seja, o modelo tem ênfase na criação de vários fatos atendendo todos os assuntos requeridos como a produção científica e tecnológica por setor econômico, áreas de conhecimento, as linhas de pesquisa também por setor econômico, área de conhecimento e outros.

3.5.2 *BACK-END*

Assim como os outros *DM*s, também o modelo de dados do *DM* de Currículo foi concebido sobre a plataforma do SGBD Oracle® e seu processo de *ETL* também foi desenvolvido usando linguagem de programação JAVA®. Uma vez ao dia os dados origem que fazem a composição do *DM* são extraídos do banco de currículos Lattes assim como a cópia integral dos arquivos XML contendo os currículos que estão no sistema de arquivos da infra-estrutura, sendo depositados em um repositório específico para o *DM* de Currículos. Para gerenciar a carga dos currículos, é necessário verificar uma tabela que mantém quais currículos são inseridos ou alterados na base de currículos Lattes diariamente, assim esses mesmos currículos serão carregados no *DM* e em caso de problemas durante o processo são gerados logs em tabelas de erros que são verificadas diariamente onde os currículos que apresentaram problemas são marcados para serem novamente carregados no próxima ciclo de execução do processo. Além da análise das consultas utilizadas pelos sistemas de consulta e manipulação dos dados do *DM* criando-se índices de auxílio a performance, também foram criadas visões baseadas em dimensões com dados filtrados para evitar as junções de consultas diretas entre fatos e dimensões que possuam uma grande quantidade de registros.

3.5.3 *FRONT-END*

Como *Front-end* foram criados sites que subsidiam a busca e análise de informações diretamente ligadas aos currículos de pesquisadores e estudantes do país, onde no endereço http://buscatextual.cnpq.br/buscatextual/index.jsp pode realizar buscas textuais por currículos ou produções, levantando os pesquisadores mais relevantes em suas áreas de atuação ou as produções mais relevantes. Também é possível determinar demograficamente indicadores da pesquisa e produção nas instituições como mostra a Figura 18, através do site de Demografia Curricular (http://demografia.cnpq.br/dmcurriculo/), e por meio do Lattes Egressos (http://egressos.cnpq.br/lattesegressos/) pode-se apresentar diversos indicadores demográficos sobre egressos conhecendo assim as distribuições geográfica, demográfica e profissional (por área ou instituição de atuação) dos egressos de cursos

de nível superior, geradas a partir de diferentes cruzamentos de dados curriculares, como mostra a Figura 19, o qual foi desenvolvido utilizando o modelo proposto por Bovo [2004], que traduz fontes de informação em um formato padrão de representação de relacionamentos entre elementos do domínio do problema, de forma a viabilizar a extração de conhecimento por meio da aplicação de *Link Analysis* e Teoria dos Grafos.

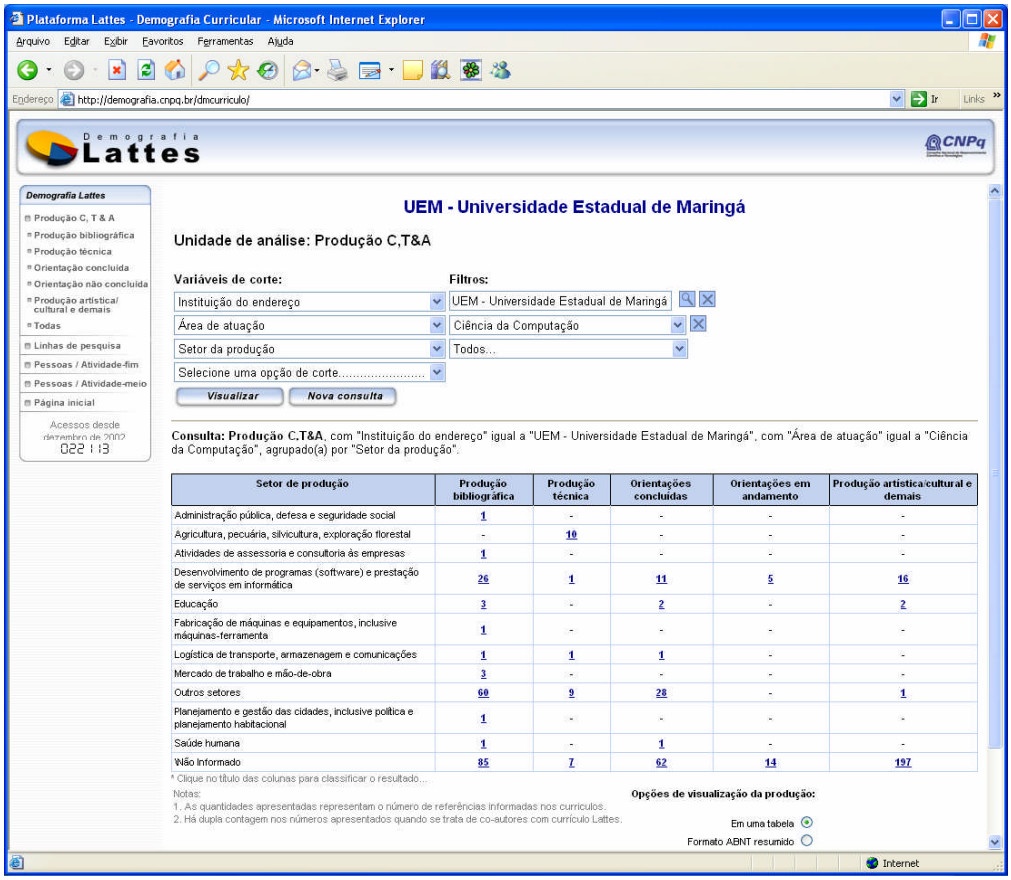


Figura 18 – Demografia Curricular

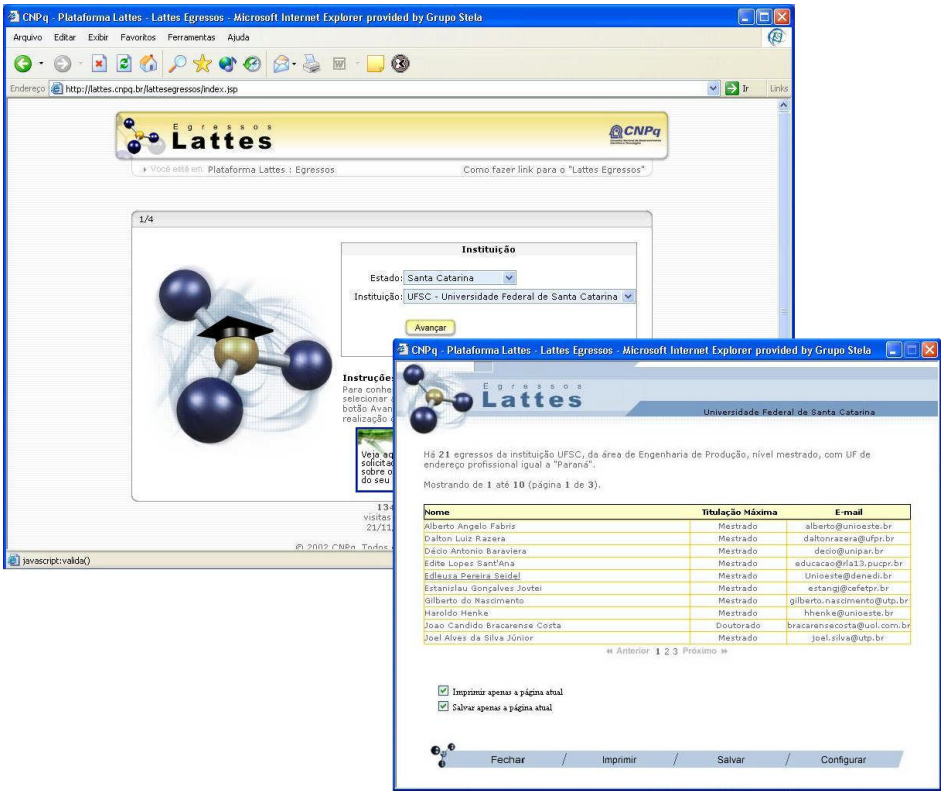


Figura 19 – Lattes Egressos

3.6 CONCLUSÃO DO CAPÍTULO

Neste capítulo foi apresentado o *DW* do CNPq onde foi abordado a arquitetura de *DM*s que o formam, os modelos adotados em cada um, seus processos de extração, limpeza e carga dos dados, formas de otimização e ferramentas *OLAP* para acessar informação disponibilizada, com o intuito de fundamentar o estudo comparativo realizado para levantar indicadores de relevância na construção de *DW*s que serão avaliados no próximo capítulo.

4 O ESTUDO REALIZADO

4.1 INTRODUÇÃO

A partir da definição do escopo do projeto de *DW* e o levantamento de seus requisitos, todo trabalho de construção de um *DW* passa a ser direcionado para a construção de seu modelo de dados, das ferramentas de *Back-End* e de *Front-End*. Tais fases são constituídas de passos que devem ser minuciosamente observados e realizados de maneira a tornar a próxima fase viável.

Os passos realizados em cada fase fornecem a base para que o *DW* seja desenvolvido a fim de torne-se uma ferramenta de extração de conhecimento eficiente, assim a aplicação das melhores técnicas em cada passo ditam o resultado final de toda a construção, tornando-se estes grandes indicadores a serem avaliados. No processo de modelagem devem ser avaliados os passos ligados a redundância de informação, volatilidade dos dados, temporalidade, flexibilidade de ajustes do modelo, granularidade e a agregação estes indicadores determinam a qualidade da informação a ser disponibilizada e a reutilização do modelo em outros projetos de *DW*.

Com relação ao processo de *Back-End*, o processo de carga dos dados, o volume de dados e índices, a indexação elaborada e a criação de agregados, são os passos a serem observados pois através destes é possível todo o planejamento dos equipamentos que serão necessários para dar suporte ao *DW*.

Já o processo de *Front-End* leva em consideração a redução da complexidade dos dados, a utilização de ferramentas *OLAP*, a performance e complexidade das consultas realizadas nos dados pois é necessário que o usuário final encontre rapidamente a informação que está buscando de maneira legível.

A seguir, serão explanados cada um dos indicadores acima para cada um dos *DMs* utilizados no estudo, em uma análise comparativa, afim de demonstrar como cada

indicador foi implementado para cada *DM* de maneira a ser mais eficiente e mostrar a sua importância em todo o processo de construção do *DW* do CNPq.

4.2 MODELAGEM

Para garantir a qualidade dos dados relacionada a uma estrutura simples e eficiente de dados, faz-se necessário o estudo de modelos de dados para o desenvolvimento de um *DW*. A modelagem de dados de um *DW* é baseada na modelagem dimensional também conhecida como esquema estrela, pois tende a manter um fato ligado a dimensões, contudo a complexidade dos assuntos tratados podem vir a influenciar o modelo tradicional. A necessidade de criar um modelo que não apresente redundância de informação, volatilidade dos dados, mantenha a temporalidade da informação, a flexibilidade de ajustes no modelo, um grau de granularidade que atenda com tempo aceitável e dados corretos a informação que está sendo buscada, agregações para aumentar a performance de resposta das consultas e a possibilidade de reutilização do modelo em novos projetos de *DW*, fornecem parâmetros para a criação de um modelo eficiente.

O *DM* de Fomento é um caso tradicional de um modelo dimensional analítico para *DW*, pois aborda apenas o assunto de pagamento de bolsas pelo CNPq, representado pelo fato pagamento que ligado a várias dimensões tais como geografia, instituição, área de conhecimento, e outras, seguindo o formato estrela, fornece a possibilidade de obter várias análises sobre o assunto.

O modelo do *DM* de Grupo de Pesquisa também mantém a característica de um modelo dimensional analítico, tratando o assunto censo dos grupos de pesquisas, mas este tem uma particularidade que é a utilização de “dimensões ponte” para auxiliar a necessidade de relacionamentos 1 para n encontrados entre alguns fatos e dimensões, como por exemplo o fato empresa por grupo, onde cada empresa pode manter vários setores de atividade econômica foi criada a dimensão de grupo de setores de atividade econômica que pode ser utilizado posteriormente em outras uniões de empresa por grupo.

O modelo do *DM* de Currículos foi concebido como um modelo agregado para atender os vários assuntos encontrados na fase de levantamento de requisitos de maneira a não gerar uma quantidade imensa de registros nos fatos e auxiliar na performance das

consultas, como exemplo pode-se citar o fato de produção por área de conhecimento e o fato de produção por setor econômico.

Será explanado a seguir os indicadores ligados diretamente a criação do modelo de dados de um *DW*.

4.2.1 Redundância de informação

Em *data warehousing*, verifica-se o armazenamento redundante de informações, havendo necessidade de manutenção do controle desta redundância. A redundância produz um aumento do volume dos dados armazenados. Um grande número de acessos vai requerer que as funções do sistema apresentem alto desempenho. Para tanto, é importante que os sistemas de bancos de dados sejam construídos de forma a garantir às aplicações o padrão de desempenho exigido por elas. São oferecidos os bancos de dados paralelos e, pelas características que apresentam, os bancos de dados paralelos mostram-se adequados para as atividades *data warehousing*, Garcia [2000].

Mas um equilíbrio entre uma redundância aceitável onde dados oriundos vindos dos sistemas operativos são replicados em poucos dimensões ou fatos e uma redundância onde esses mesmos dados se repetem em várias dimensões ou fatos determina a custo de *hardware* de armazenamento de um *DW*.

Contudo na modelagem dos *DMs* do CNPq o compartilhamento de dimensões entre os modelos foi sempre levado em consideração, onde a exemplo a dimensão que mantém a área de conhecimento dos pesquisadores, a dimensão que armazena os dados pessoais de cada pessoa ou a dimensão onde estão todas as instituições acadêmicas, de fomento ou pesquisa, são acessadas tanto pelo *DM* de Fomento, de Grupo de Pesquisa e de Currículos. O único modelo que mantém uma maior redundância de informações relacionadas aos fatos é o modelo do *DM* de Currículo, este contém vários fatos de produções científicas e tecnológicas que se dividem por setor e área de conhecimento, setor e grande área de conhecimento, e outros, assim como linha de pesquisa por setor e área de conhecimento, setor por grande área de conhecimento, mas estes são exemplos de redundância aceita para um ambiente de *DW* pois além de fornecer informações essenciais

sobre a produção científica e as linhas de pesquisa no país tornam a respostas das consultas relacionadas mais rápidas.

4.2.2 Volatilidade dos dados

Para que o *DW* possa ser consistente e integro no decorrer de todo o tempo e fornecer comparações confiáveis entre períodos distintos é necessário que as informações nele contido não devam ser atualizadas, mas em caso de mudança, inseridas novamente como novos registros a fim de manter a consistência dos dados ao longo da sua evolução.

Para isso todas as dimensões do *DM* de Fomento mantém chaves artificiais, onde a cada carga incremental de dados, regras de verificação de alteração são aplicadas e se uma alteração é detectada novos registros serão inclusos nas dimensões de modificação lenta, já em relação ao fato de pagamento, este é inserido novos registros pertinentes aos pagamentos das bolsas mensalmente mantendo a integridade dos dados em seus períodos distintos.

O *DM* de Grupo de Pesquisa também mantém chaves artificiais para que se possa incluir registros novos nas dimensões de modificação lenta em caso de atualização da origem ou mudança do ano do censo, mas a cada atualização de um grupo, seus dados relacionados ao censo atual são apagados do *DM* e novamente carregados.

O *DM* de Currículo apresenta a mesma peculiaridade que o *DM* de Grupo de Pesquisa neste indicador, pois se constatado pelas regras de verificação que um determinado indivíduo mantém as características básicas de suas informações pessoais, todo seu currículo é inicialmente apagado do *DM* e novamente carregado, caso contrário, um novo registro é gerado para esse indivíduo em cada uma das tabelas de dimensões e fatos.

4.2.3 Temporalidade

A possibilidade de traçar uma análise histórica e comparativa entre os fatos ocorridos ou retratar a situação que estamos analisando, num determinado ponto do tempo

é uma característica pertinente ao ambiente de *DW*, tornando assim a temporalidade incontestavelmente importante, pois é sobre a variabilidade dos dados em torno dela que é tomada as decisões estratégicas.

O modelo do *DM* de Fomento provê a dimensão de tempo, que associada ao fato de pagamento de bolsas possibilita análises históricas sobre o fomento do CNPq.

Já o modelo do *DM* de Grupo de Pesquisa mantém a dimensão de censo que informa os anos dos censos realizados que cada grupo de pesquisa foi incluído.

O modelo do *DM* de Currículo não possui uma dimensão de tempo ou algo parecido, mas várias análises históricas podem ser realizadas através do ano da produção científica ou tecnológica nos fatos de produções, ou pelo período nos fatos de linha de pesquisa. Análises sobre formação acadêmica podem ser realizadas sobre o atributo de ano de início e término da formação, ou ano de obtenção do título na dimensão de formação, assim como análises temporais das atividades acadêmicas realizadas observando os atributos de ano de início e término das atividades na dimensão de atividades.

4.2.4 Flexibilidade de ajuste

Na implementação de um *DW* o modelo deve permitir que as informações nele contido devam servir de origem para qualquer consulta gerencial que possa ser desejada, tanto pelos diversos setores da corporação quanto pelo mais alto nível hierárquico de gestão, assim a inclusão de ajustes no modelo deve exigir o menor impacto possível para que o modelo continue a atender seus objetivos já levantados e os novos que foram gerados posteriormente.

O *DM* de Fomento é o modelo que tem maior facilidade de ajuste por ser concebido como um tradicional modelo dimensional analítico oferece a flexibilidade de inclusão de novos atributos ou dimensões que consequentemente permanecerá fornecendo os resultados requeridos das unidades de análise já criadas e será capaz de responder as novas análises projetadas.

Também o modelo de *DM* de Grupo de Pesquisa poderá sofrer ajustes para abrigar novas unidades de análise na sua estrutura devido a este modelo também ser um modelo dimensional analítico, apenas deverá ser observado se o ajuste requerido influenciará ou dependerá de criação de novas "dimensões ponte".

Contudo o modelo do *DM* de Currículos se apresenta como um modelo mais "engessado", pois ajustes derivados de criação de novas unidades de análise irão provocar a criação de novos fatos agregados para atender as novas consultas, podendo ser necessário a busca de novos dados dos sistemas operativos para formar novas dimensões que venham suportar os novos fatos.

4.2.5 Granularidade

Definir o nível de detalhamento de um *DW* exige que este seja equilibrado, pois se a granularidade for muito baixa, ou seja, um nível de detalhamento alto, fará com que eleve-se a quantidade de registros nos fatos, implicando assim diretamente na performance do sistema e consumo de armazenamento, mas fornecendo a possibilidade de atender todo tipo de consulta realizada aos dados, caso contrário um nível mais baixo impede que sumarizações e agregações sejam realizadas pois os fatos já estarão resumidos, mas tornando o consumo de espaço físico de armazenamento muito baixo e elevando consideravelmente a performance do sistema.

Uma característica inerente a todos os três modelos de *DM* é a possibilidade de que todos podem chegar a nível de detalhamento de pessoa, pois suas unidades de análise sempre tem como parte do conjunto de atributos a identificação da pessoa, por exemplo no *DM* de Fomento o fato pagamento de bolsas contem o atributo pessoa, assim como no *DM* de Grupos de Pesquisa o fato de pessoa por grupo, também possui o atributo pessoa e finalmente no *DM* de Currículos os fatos de produção científica e tecnológica e seus afins contem o atributo pessoa, ou os fatos de linhas de pesquisa e afins também o possuem.

4.2.6 Agregados

Definir as dimensões com níveis de agregação e fatos em um *DW* influência diretamente no *Front-End*, pois operações de "*drill-down*" ou "*roll-up*" realizadas por ferramentas *OLAP* só serão possíveis se assim o modelo for concebido.

Ao fazer uso de agregações, deve-se ter em mente que se trata da ferramenta mais poderosa disponível em um *data warehouse* em termos de controle de desempenho. Uma agregação nada mais é do que uma tabela de fatos que sumariza outros fatos de nível mais baixo.As questões de performance são mais críticas nas aplicações de *data warehouse*. O fato de que um *DW* é um sistema de suporte a decisões de nível gerencial, faz surgir a necessidade de um DBA mais atento a satisfação do usuário final. Aspectos como, por exemplo, o custo da hora de trabalho de um executivo e o custo de armazenamento e carga dos dados, devem ser analisados e monitorados constantemente, com o objetivo de reavaliar a necessidade de agregação e índices.

Existem duas técnicas de agregação: criação de novas tabelas de fatos agregadas e a criação de novos campos de indicação de nível de agregação nas tabelas dimensões. No primeiro caso, exige-se a criação de chaves artificiais em cada uma das dimensões que está sendo agregada. Já no caso da criação de campos de indicação do nível de agregação, os fatos agregados serão incluídos na própria tabela de fatos original. Observa-se que esta segunda opção não é a mais recomendada por gerar problemas de dupla contagem.

É necessário Controlar a explosão das agregações, pois, as agregações podem causar uma explosão do tamanho do *data warehouse*, no caso de processos comerciais. Uma das formas eficientes de controlar esta explosão é garantir que cada agregação sumarize mais de 10 ou 20 itens. Também deve-se balancear os índices entre as agregações, pois, a combinação do uso Índices em agregados é bastante interessante, pois tira da tabela de fatos o peso de carregar todos os índices, distribuídos por entre as tabelas de fatos agregadas.

Como existe um certo compartilhamento de dimensões entre os *DM*s no geral as dimensões de pessoa e instituição fornecem agregações, como por exemplo, agrupamento por sexo e possibilidade de descer ao nível de nome da pessoa na dimensão pessoa, ou o agrupamento por região geográfica e a possibilidade de descer ao nível de nome da

instituição na dimensão de instituições. No *DM* de Fomento encontra-se agregação nas dimensões de tempo onde o agrupamento por ano pode descer até o nível de dia, ou na dimensão de cursos acadêmicos onde o agrupamento por instituições pode descer ao nível de nome do curso, assim como outras dimensões que também oferecem outros tipos de agregação. O *DM* de Grupo de Pesquisa oferece na dimensão de setor da atividade econômica do grupo de pesquisa possibilidade de agregação, onde a partir do agrupamento dos níveis de cada setor pode se chegar ou setor específico ou na dimensão de empresas, onde a partir do agrupamento de região geográfica pode-se chegar ao nome da empresa. Já o *DM* de Currículos mantém a dimensão de produções científicas e tecnológicas de cada currículo que pode agrupar as produções por tipo o subtipo de produção até chegar a produção de uma pessoa, ou a dimensão de área de formação que pode agrupar por grande área, área chegando ao nível de pessoa.

4.2.7 Reutilização de modelos

A possibilidade de se reaplicar um modelo sobre novos *DW* que irão abordar o mesmo assunto, torna este modelo apto a ser um *template* de modelo de dados reutilizável. Tanto o modelo do *DM* de Fomento como o de Grupo de Pesquisa fornecem a possibilidade de serem reutilizados em novos *DM*s que abordem assuntos semelhantes, por se tratarem de modelos dimensionais tradicionais fundados sobre o esquema estrela fornecendo a possibilidade destes novos *DM*s se reutilizarem de suas dimensões e fatos, porém o modelo do *DM* de Currículos não fornece tal facilidade pois trata muito especificamente os assuntos já determinados no seu levantamento de requisitos, sendo este modelo um modelo agregado que fornece o resultado de cada assunto pronto nos seus fatos praticamente sem a necessidade de utilização de junção entre dimensões e fatos, assim a reutilização do modelo só seria possível se fosse para tratar um novo *DM* que tivesse praticamente todas as características do *DM* de Currículos.

4.3 BACK-END

Todo o processo de extração dos dados dos sistemas operativos, sua limpeza e a carga destes dados para o *DW*, o dimensionamento do espaço que será utilizado para os

dados e índices que serão gerados além dos agregados projetos para aumentar a performance de resposta do *DW*, visam a qualidade da informação e a performance do *DW* e são fisicamente concebidos na fase de *Back-End*, justificando a necessidade da avaliação dos mesmos.

4.3.1 Processo de carga (Performance, eficiência)

A extração, limpeza e carga dos dados é considerada a operação de maior custo para todo o *DW*, buscar os dados nas fontes origens, deixá-los preparados utilizando regras de validação e transformação e por fim povoar as dimensões e sumarizar os fatos, pode demandar alto poder de processamento.

No *DM* de Fomento foi utilizada a própria linguagem de programação do SGBD Oracle®, criando-se um pacote que contém as funções e procedimentos necessários a realização de todo o processo de carga dos dados, fazendo com que o processo se torne mais performático, outra condição que implica na performance é o fato de que os dados origens estão todos no mesmo SGBD, evitando assim o tráfego de rede dos dados e a necessidade de se utilizar de outras ferramentas para acessar SGBDs de outros fabricantes. O processo se baseia na busca dos dados origens, fazendo um mapeamento do que está sendo buscado no operacional com o que está carregado no *DW* em uma área de estagiamento, para que se possa manter a integridade dos dados já carregados e então carregar as dimensões e gerar o fato de pagamento de bolsas. Este processo é realizado apenas uma vez a cada mês, fazendo com que a utilização do processamento de *hardware* seja pequena e de curta duração.

O processo de carga do *DM* de Grupo de Pesquisa foi todo feito em linguagem de programação JAVA®, não compilada internamente no SGBD Oracle®, e consiste na busca dos dados necessários para a carga do *DM* no ambiente operacional do Diretório de Pesquisadores. Assim como o *DM* de Fomento os dados origem para o *DM* de Grupo de Pesquisa também estão no mesmo SGBD ajudando na performance do processo de carga que consiste na verificação na origem dos dados para levantar quais grupos sofreram atualização desde a última carga e então apagá-los da *DM* e carregá-los novamente, o que tende a consumir mais tempo de processamento pois exige do SGBD a validação dos

índices e relacionamentos ligados as tabelas que sofrerão o *delete,* se o grupo é novo este é inserido no *DM* e por fim são alimentados os fatos.

O *DM* de Currículos tem um processo de carga muito semelhante ao do *DM* de Grupo de Pesquisa, este também foi feito em linguagem de programação JAVA®, não compilada internamente no SGBD Oracle®, e a aplicação diariamente verifica os currículos que foram atualizados ou inseridos na base operacional e então carrega-os para o *DM*, onde caso um currículo passe pelas regras de verificação como sendo um currículo novo este é inserido, caso contrário, as informações pertinentes ao currículo são apagadas e então recarregadas nas dimensões, para então serem gerados os fatos, contudo este *DM* também mantém o *XML* de cada currículo carregado na base e para se carregar este *XML* é necessário fazer a leitura no sistema de arquivo do sistema operacional, pois a geração desse *XML* é feito por outro sistema e depositado no sistema de arquivo, o que aumenta o processo de *input*/*output* de hardware.

4.3.2 Volume de dados e índices

O volume de dados e dos índices estão diretamente relacionados a granularidade do *DW* e a complexidade das consultas realizadas pelos usuários. Como visto anteriormente quanto maior o nível de detalhamento maior será o espaço físico ocupado em *hardware,* consequentemente aumentando o espaço ocupado também pelos índices, e quanto maior for complexo as consultas dos usuários mais índices terão de ser construídos para diminuir o tempo de resposta das mesmas.

O *DM* de Fomento tende a manter um número permanente de acréscimo de bytes ocupados, pois mensalmente são lançados os pagamentos das bolsas a bolsistas, por exemplo, se pegarmos o ano de 2001 entre janeiro e dezembro do mesmo ano a um incremento médio mensal de 46.800 registros no fato pagamento de bolsas em compensação as dimensões são na sua maioria apenas atualizadas e com uma carga de inserção praticamente nula. As consultas realizadas sobre o *DM* de Fomento foram analisadas e além da criação dos índices de chave primária e estrangeira nas tabelas de dimensão e fato, também foram criados índices por um atributo ou por um grupo de

atributos afim de atender todas as consultas, mas apesar disso o volume de espaço ocupado em *hardware* e na ordem do dobro do espaço utilizado pelos dados.

O *DM* de Grupo de Pesquisa assim como o *DM* de Fomento mantém um crescimento homogêneo com relação aos grupos de Pesquisa do país, levando em consideração o censo de 2000 até o atual, a média de grupos por censo é cerca de 16479 grupos, contudo este *DM* mantém os dados de cada grupo em formato *XML* (linguagem de marcação de dados), tal estrutura ocupa um espaço considerável do SGBD em *hardware*. Os índices criados para as chaves primárias das dimensões e chaves estrangeiras dos fatos atendem parte das consultas realizadas, mas foram analisadas as consultas realizadas no *DM* para se criar outros índices que atendessem aquelas que não estavam se utilizando dos índices já criados, e mesmo com a criação de novos índices o espaço ocupado em *hardware* foi de ¼ do espaço ocupado pelos dados.

Dentre os três *DMs* o de Currículos é o que mais consome espaço em hardware, sua média diária é de cerca de 4.000 currículos por dia (entre currículos novos e atualizados) e como o *DM* de Grupo de Pesquisa, este também mantém os dados de cada currículo em estrutura *XML*. Se comparado ao volume total ocupado em hardware de dados, o volume dos índices é pequeno, pois hoje estimasse um espaço ocupado de cerca de 50gb de dados e apenas um consumo de apenas 6% desse valor em índices criados no SGBD.

4.3.3 Indexação

Um dos mais relevantes fatores para a performance de um *DW* é a análise e criação de índices para auxiliar as consultas geradas pelos usuários, este recurso diminui consideravelmente a quantidade de leitura e escrita no hardware de armazenamento físico do *DW*, e retorna em segundos seleções complexas em enormes bases de dados. Utilizando o método de indexação padrão do Oracle®, índice b-tree (objeto estruturado de árvore balanceada), foram criados os índices de chave primária e estrangeira para o *DM* de Fomento e não houve a necessidade de criação de índices de estruturas mais complexas, apenas foram analisadas as consultas realizadas pelas ferramentas OLAP e gerados índices que atendessem de maneira a aumentar a performance de resposta. Porém para atender as consultas realizadas pela busca textual do fomento, foi usado o projeto *Apache Jakarta*

Lucene [APACHE SOFTWARE FOUNDATION, 1997] que é um motor de busca textual de alta performance desenvolvido totalmente em JAVA® e que gera no próprio sistema de arquivos do sistema operacional uma estrutura de indexação. Este método de indexação mostrou-se muito mais eficiente e rápido do que se tivesse sido criada uma estrutura de indexação proprietária do Oracle® para consulta textual chamada de "*Context*".

Também o *DM* de Grupo de Pesquisa se utiliza da indexação padrão do Oracle®, e para auxiliar as consultas das ferramentas *OLAP*, além dos índices de chave primária e chave estrangeira gerados para as dimensões e fatos foram criados novos índices que forneceram melhor performance nas consultas realizadas. Para atender a busca textual *DM* Grupo de Pesquisa também foi utilizado a ferramenta *Lucene*.

Da mesma maneira que o *DM* de Fomento o *DM* de Currículos se utiliza de indexação padrão do Oracle®, e também foi baseado a construção dos índices nas chaves primárias e estrangeiras das dimensões e fatos, e para ajudar na performance das consultas feitas ao *DM*, foram criados índices que atendem a uma ou mais consultas. Este *DM* também se beneficia da performance fornecida pela ferramenta *Lucene* para atender as buscas textuais que substituiu a indexação *Context* do Oracle® na primeira versão do sistema de busca textual realizado nos XMLs dos currículos devido a sua baixa performance em uma base de dados de grande volume.

4.3.4 Criação de agregados

Uma característica importante do *DW* é apresentar dados agregados, buscando primeiramente aumentar o desempenho das consultas realizadas por seus usuários. As agregações garantem ainda uma redução no tempo de resposta de processamento e uma redução de espaço de armazenamento. Infelizmente o uso da agregação dos dados no *DW* pode acabar reduzindo a capacidade ou funcionalidade do *DW*, ocasionando a perda de detalhes.

Para o *DM* de Fomento foi utilizada a técnica de criação de visões materializadas do SGBD Oracle®, que consiste em criar fisicamente tabelas que podem ser agregações de fatos. Dependendo da consulta sumarizada requerida pelo usuário, o SGBD verifica se deve buscar os dados sumarizando diretamente na tabela de fato ou se é mais performático

redirecionar a consulta para uma visão materializada que já mantém os dados sumarizados. Uma fator importante é que até mesmo nestas visões materializadas pode-se aplicar a construção de índices, o que torna a resposta da consulta ainda mais rápida, como exemplos pode-se citar a visão materializada que agrega o pagamento de bolsas por área de conhecimento, região geográfica, sexo e faixa etária, ou a visão materializada que agrega o pagamento de auxílios à pesquisa por área de conhecimento, região geográfica, sexo e faixa etária, e também a visão materializada que agrega o pagamento de investimentos por fundos setoriais, agrupados por região geográfica e área de conhecimento. Devido ao baixo volume de registros no *DM* de Grupo de Pesquisa, não houve a necessidade de criação de agregados, sendo a análise e criação de índices suficiente para atender com rapidez as consultas realizadas ao *DM*.

O *DM* de Currículos é o que contém a maior utilização de agregados, como por exemplo o fato que mantém as linhas de pesquisa por área de conhecimento tem sua versão agregada no fato de linhas de pesquisa por grande área de conhecimento, ou o fato das produções científicas e tecnológicas agrupado por setor de atividade e área de conhecimento tem os dados agrupado por setor de atividade e grande área de conhecimento.

4.4 *FRONT-END*

Como resultado visual do *DW* o *front-end* que será utilizado pelos usuários no seu cotidiano profissional, necessita disponibilizar aos usuários as informações de maneira fácil e inteligível, além de tornar o tempo das respostas das consultas aceitável de se esperar e a possibilidade dos usuários aplicarem ferramentas OLAP sobre o *DW* a fim de poder obter mais conhecimento a partir deste repositório, tornando fundamental a análise dos quesitos, legibilidade, utilização de ferramentas OLAP e otimização das consultas realizadas para se obter assim um *front-end* capaz de responder as necessidades dos usuários do *DW*.

4.4.1 Legibilidade

Permitir fácil acesso à informação, com conteúdo intuitivo e significado das informações óbvio para os usuários é o necessário para definir a legibilidade de um modelo que é acessado por uma ferramenta de *Front-End*. Modelos dimensionais demonstram ser propícios para a legibilidade do usuário, pois com a aplicação de uma ferramenta *OLAP* sobre o modelo o usuário irá identificar com clareza as unidades de análise e poderá facilmente montar os cubos multidimensionais para obter a informação que deseja.

O modelo do *DM* de Fomento fornece grande legibilidade pois foi concebido sobre o esquema estrela que permite ao usuário identificar facilmente as unidades de análise, como instituições, região geográfica cursos ou área de conhecimento que ligados ao fato de pagamento de bolsas e auxílios fornecem valiosas informações sobre o investimento em pesquisa no país.

O modelo do *DM* de Grupo de Pesquisa oferece está legibilidade, pois também foi projetado sobre o modelo estrela, sendo possível ao usuário escolher por exemplo as dimensões de instituições, região geográfica ou área de conhecimento e uni-las ao fato de grupo de pesquisa obtendo um cubo para análise de informações sobre a evolução do censo de grupos de pesquisa no país.

O modelo do *DM* de Currículos é o que apresenta maior deficiência na legibilidade, pois foi construído totalmente focado em seus assunto e performance de resposta, apresentando uma quantidade enorme de agregados com por exemplo os fatos de produção, linhas de pesquisa e pessoa, que atende a cada assunto definido no levantamento de requisitos tornando assim o *DM* um modelo que apresenta um grau de complexidade mais elevado para ser utilizado em uma ferramenta *OLAP*, onde o usuário não tem clareza de como as informações estão organizadas.

4.4.2 Utilização de OLAP

As manipulações sobre o modelo multidimensional que permitem uma visão de negócio em diferentes perspectivas geram cubos a partir das dimensões de forma que as

consultas realizadas sobre este cubo possam visualizá-lo e manipulá-lo sob diferentes ângulos e diferentes níveis de agregação. Devido a utilização de um modelo dimensional para a geração dos *DM*s, recursos *OLAP*s puderam ser aplicados as ferramentas de *Front-End* que acessam os dados. Por exemplo, se utilizado o site de Investimentos do CNPq em CT&I, será criado visões multidimensionais em tempo de execução a partir da escolha de uma categoria de investimento, ano e agrupamento e com o resultado obtido nesse cubo é possível realizar operações de "*drill-down*" até o grão pessoa. Com a utilização do site de Diretório dos Grupos de Pesquisa no Brasil, é possível por exemplo no Plano Tabular[1] escolher uma unidade de análise e aplicar filtros que irão criar a visão multidimensional que poderá sofrer um "*drill-down*" até o grão pessoa. No site de Demografia Curricular[2] criado para acessar os dados do *DM* de Currículo através de escolha da instituição, o indicador de análise, variáveis de corte e filtros o resultado da visão multidimensional pode chegar até ao grão pessoa através de "*drill-down*".

4.4.3 Consultas (Performance, complexidade)

Para que o tempo de resposta no processo de formulação de uma consulta, seu envio e execução no banco de dados e o retorno desta ao usuário seja em tempo aceitável é necessário que o *DW* esteja pronto para fornecer essa performance. Invariavelmente a complexidade de uma consulta vai influenciar no resultado de resposta e para se evitar esta situação, o modelo utilizado e a analise da execução de cada possível consulta devem ser muito bem projetados.

No *DM* de Fomento, uma consulta que necessita trazer a quantidade de bolsas e auxílios pagos por ano e área de conhecimento é necessário a junção do fato pagamento com as dimensões tempo e área de conhecimento, mas sem a especificação do ano e de uma área de conhecimento a tendência é que se retorne um grande número de registros pois será uma consulta aplicada em uma base com milhões de registros, mas se for aplicado as técnicas do SGBD de análise de tabelas e índices de chave primaria e chave estrangeira, o resultado da consulta retornará em tempo aceitável pelo usuário, claro que se em cima da consulta ainda fossem aplicados o filtros de ano e área de conhecimento a

[1] Ferramenta que permite visualizar quantitativamente o perfil da pesquisa no Brasil.

[2] Ferramenta que visa apresentar o perfil da pesquisa e produtividade de pessoas vinculadas a uma instituição quanto a distribuição da população pesquisada.

quantidade de registros seria menor, mais índices seriam utilizados e consequentemente a resposta da consulta seria em menor tempo, contudo quanto mais unidades de análise forem requeridas maior será a complexidade do consulta pois maior será a quantidade de junções com outras dimensões, levanto ao aumento de tempo de resposta das consultas. A mesma consulta aplicada no *DM* de Grupo de Pesquisa apresentou ser menos complexa e de maior performance pois para agrupar os grupos de pesquisa por ano e área de conhecimento foi necessário apenas a junção do fato de grupos com a dimensão de área de conhecimento pois o ano de censo já está presente no próprio fato de grupos, e com a aplicação dos filtros a performance é ainda maior, mas deve-se levar em consideração a quantidade de registros da base que está apenas na ordem de milhares.

Entre os três *DM*s o *DM* de Currículo é o que apresenta o modelo de maior performance aos assuntos por ele abordado, sendo que um modelo agregado mantém vários fatos que atendem plenamente as consultas realizadas, como por exemplo o caso de busca de produções por área de conhecimento e ano da produção, é uma consulta simples realizada apenas sobre o fato de produções por área de conhecimento pois este já mantém o ano da produção e a área de conhecimento no próprio fato e mesmo a base estando na ordem de milhões de registros o tempo de resposta é muito baixo.

4.5 RESULTADOS

Com o levantamento comparativo entre os indicadores dos processos de Modelagem, *Back-End* e *Front-End*, pode-se criar uma tabela com os indicadores estudados em cada processo e observar o comportamento dos *DM*s de Fomento, Grupo de Pesquisa e Currículos, onde este comportamento pode ser visualizado na tabela 4.

Tabela 4 - Indicadores Comparativos

Indicadores / ***DMs***		***DM* Fomento**	***DM* Grupo**	***DM* Currículo**
Modelagem	Redundância de informação	baixa redundância	Baixa redundância	Nível de redundância médio
	Volatilidade dos dados	Não volátil	Volátil	Volátil
	Temporal	Mantém temporalidade dos dados	Mantém temporalidade dos dados	Mantém temporalidade dos dados
	Flexibilidade de ajuste	Facilmente ajustável	Facilmente ajustável, mas se o ajuste influenciar as "dimensões pontes" pode se tornar um pouco mais complexo	Adaptabilidade complexa, pois deverão ser criados novos fatos agregados para suprir as novas unidades de análise.
	Granularidade	Nível de detalhamento de pessoa	Nível de detalhamento de pessoa	Nível de detalhamento de pessoa
	Agregados	Possui agregação feita através de visões materializadas	Possui agregação	Possui agregação feita através do processo de ETL

	Reutilização de modelos	Possível reutilização	Possível reutilização	Possibilidade de reutilização em projetos de *DW* que tratem assuntos bem semelhantes
Back-End	Processo de carga	Performance alta apesar de ser necessário a carga em algumas dimensões além dos fatos e devido a freqüência de execução, utilização de linguagem proprietária do SGBD e manipulação de dados não complexos	Performance média devido a execução diária, necessidade de apagar e recarregar dimensões e fatos, utilização de linguagem JAVA® não compilada diretamente no SGBD e manipulação de estruturas XMLs	Performance média devido a execução diária, necessidade de apagar e recarregar dados apenas de fatos, utilização de linguagem JAVA® não compilada diretamente no SGBD e manipulação de estruturas XMLs
	Volume de dados e índices	Baixo volume de dados e índices devido a baixa inclusão mensal de novos registros	Baixo volume de índices mas Médio volume de dados devido a utilização de estrutura XML para armazenar do currículo do grupo	Baixo volume de índices mas Grande volume de dados devido a utilização de estrutura XML para armazenar do currículo da pessoa com inclusão diária de registros
	Indexação	Atende rapidamente as consultas realizadas, além de utilização do Lucene para a busca textual	Atende rapidamente as consultas realizadas, além de utilização do Lucene para a busca textual	Atende rapidamente as consultas realizadas, além de utilização do Lucene para a busca textual

	Criação de agregados	Utilização de visões materializadas	Sem agregações devido a pouca quantidade de registros	Utilização de fatos agregados
Front-End	Legibilidade	Legível para o usuário	Legível para o usuário	Legibilidade complexa para o usuário
	Utilização de *OLAP*	Fácil criação de ferramenta *OLAP* que gera visões multidimensionais devido a estrutura dimensional do modelo	Fácil criação de ferramenta *OLAP* que gera visões multidimensionais devido a estrutura dimensional do modelo	Dificuldade na criação de ferramenta *OLAP* devido o modelo ser agregado, dificultando visões multidimensionais
	Consultas	Com boa performance e complexidade média	Com boa performance e baixa complexidade	Com alta performance e baixa complexidade

4.6 CONCLUSÃO DO CAPÍTULO

Neste capítulo foram analisados indicadores dos processos de Modelagem, *Back-End* e *Front-End* do *DM*s de Fomento, Grupos de Pesquisa e Currículos com o objetivo de explanar o comportamento de cada indicador no três *DM*s e relatar a importância de cada um no processo de *data warehousing*.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou através da avaliação das diferentes abordagens empregadas no processo de desenvolvimento de uma arquitetura de *data warehouse*, pode apresentar as características pertinentes aos processos de modelagem de dados, *back-end* e *front-end*, dos *DM* de Fomento, *DM* de Grupos de Pesquisa e *DM* de Currículos pertencentes ao *data warehouse* do CNPq.

Utilizando os elementos teóricos relacionados ao processo de *data warehousing*, pode-se levantar os indicadores de maior relevância e avaliar o comportamento de cada um aplicado a cada *DM* do CNPq podendo assim formular um quadro comparativo com a finalidade de identificar as vantagens de cada abordagem.

Para o processo de modelagem dos dados os três *Data Marts* atendem a maior parte dos indicadores utilizados no estudo comparativo, fornecendo um baixa redundância ou uma redundância aceitável em ambientes de *DW*s, apresentando volatilidade nos *DM*s de Grupos de Pesquisa e Currículo e mantendo a temporalidade da informação inserida, um nível de granularidade e agregação dos dados que fornece possibilidade de responder as consultas requeridas pelos usuários, e demonstrando apenas que o *Data Mart* de Currículos possui um grau de complexidade que o torna pouco flexível a ajustes e reutilização de seu modelo por outros *Data Warehouses*.

No processo de *back*-end pode observar que o *DM* de Fomento tem um ganho de performance sobre os outros *DM*s devido a sua execução ser mensal e ser um processo que se utiliza da linguagem proprietária do SGBD além de não necessitar manusear estruturas *XML*s. Os três *DM*s apresentam indexação proprietária que eleva a performance das consultas e utilização de indexação de alta performance para buscas textuais e possuem um baixo volume de ocupação de espaço dos índices, porém o *DM* de Currículos é um grande consumidor de espaço físico devido a grande quantidade de dados que são introduzidos diariamente. Os *DM*s de Fomento e Currículos criaram agregações para responderem com performance as consultas, com exceção do *DM* de Grupo que mantém uma quantidade de registros que são atendidas satisfatoriamente apenas pela utilização dos fatos.

No *front-end* apesar do *DM* de Currículos ter legibilidade mais complexa para o usuário, oferece maior performance na respostas das consultas devido sua pouca complexidade de junções entre dimensões e fatos, por manter vários fatos agregados que atendem sozinhos essas consultas e assim como o *DM* de Fomento e Grupos de Pesquisa atendeu as necessidades para a criação e aplicação de ferramentas *OLAP*.

Assim a escolha e aplicação de cada um dos modelos dos *DM*s possibilitou que, a arquitetura de *DW* da Plataforma Lattes, pudesse integrar e disponibilizar as informações que mapeiam a atividade científica e tecnológica do País. Os indicativos extraídos da Plataforma, quer sejam através dos instrumentos de consultas disponibilizados nos vários sites da Plataforma ou pelos sistemas de informação construídos para suportar análises dos técnicos do CNPq, têm possibilitado maior agilidade e transparência da Agência. Além disso, a integração dos dados das instituições, dos grupos de pesquisa, dos pesquisadores e bolsistas possibilitou um acesso rápido a informações que até então não eram consideradas na gestão de C&T.

Desta maneira, torna-se possível buscar uma maior eficiência nos processos ligados à gestão de C&T a fim de melhorar a qualidade dos serviços prestados pelo CNPq e alcançar a total transparência das ações do governo como agência de fomento à pesquisa indo ao encontro da e-governança.

Para trabalhos futuros sugere-se o estudo da viabilização da aplicação de unidades métricas em cada um dos indicadores a fim de torná-los quantitativos e de fácil mensuração na sua aplicação em novos projetos de *data warehouse*.

REFERÊNCIAS BIBLIOGRÁFICAS

APACHE SOFTWARE FOUNDATION, Apache Lucene. Disponível em: <http://lucene.apache.org/java/docs/>. Acesso em: 15 nov. 2005.

BOVO, B. A., Um método de tradução de fontes de informação em um formato padrão que viabilize a extração de conhecimento por meio de Link Analysis e Teoria dos Grafos. Florianópolis, 2004. Dissertação (Mestrado em Engenharia de Produção) – Engenharia de Produção e Sistemas, UFSC.

BUSINESSINTELLIGENCE.COM. Forrester Research Finds Enterprise IT Spending Intentions Show Modest Growth For 2005. Disponível em: <http://www.businessintelligence.com/ex/asp/id.850/xe/binewsdetail.htm>. Acesso em: 23 dez. 2004.

CAMPOS, Maria Luiza, FILHO, A.V. Rocha. Data Warehouse. In: XVII Congresso da Sociedade Brasileira de Computação – XVI Jornada de Atualização em Informática. Brasília, 1997, pg 221-261.

COMPUTERWORLD. AlphaBlox deal anchors IBM's BI strategy. Disponível em: <http://www.computerworld.com.au/index.php/id;807272035;fp;16;fpid;0>. Acesso em: 12 out. 2004**.**

CNPq. Conheça o CNPq, Apresentação. Disponível em: < http://www.cnpq.br/sobrecnpq/index_novo.htm>. Acesso em: 11 out 2005.

COMPUTERWORLD. The Top 10 Critical Challenges for Business Intelligence Success. 30 jun. 2003. Disponível em: <http://www.computerworld.com/services/ whitepapers/story/0,4793,82630,00.html>. Acesso em: 10 out. 2004.

DEMAREST, M. Data Legibility in Decision Support Systems (DSS). White Paper. Decision Points Applications Inc. Oregon. 2001.

GARCIA-Molina, H.; Ullman, J. D.; Widom, J. Database System Implementation. Prentice Hall, 2000.

GARTNER GROUP. Gartner says more than 50 Percent of data warehouse projects will have limited acceptance or will be failures through 2007. Disponível em: <http://www.gartner.com/press_releases/asset_121817_11.html>. Acesso em: 14 ago. 2004.

GONZAGA, T. S., Uma Metodologia para o Desenvolvimento de Instrumentos de Análise Multidimensional da Informação em Projetos de Governo Eletrônico Voltado ao Cidadão. Florianópolis, 2005. Dissertação (Mestrado em Engenharia de Produção) - Engenharia de Produção e Sistemas, UFSC.

HACKNEY, Douglas. *Data Warehouse* Delivery: Who are You? Part I. DM Review Magazine, v. 8, n. 2, 1998.

HUBER, J. et al. Integrating Data Warehouses versus Building a Federated Data Warehouse: A Comparision. In: Proceedings of the 3rd International Conf. on Data Warehousing and Knowledge Discovery, Munich, Germany, 2001. Disponível em: http://www.scch.at/servlet/resource.ResourceLoader?id=120.

IGARASHI, W.;TODESCO, J.; SELL, D.; PACHECO, R.C.S.;MARQUES A.,A. Arquitetura de *Data Warehouse* da Plataforma Lattes. In: Conferência Sul-Americana em Ciência e Tecnologia Aplicada ao Governo Eletrônico, Florianópolis, 2004.

INMON, W. H.., Como Construir o *Data Warehouse*. Campus, Rio de Janeiro, 1997.

KIMBALL, R. *Data Warehouse* Toolkit: Técnicas para Construção de *Data Warehouse* Dimensionais. São Paulo: Makron Books, 1998.

MACHADO, Felipe Nery Rodrigues. Projeto de *Data Warehouse* – Uma visão Multidimensional. São Paulo: Érica, 2000.

PACHECO, R.C.S. Uma Metodologia de Desenvolvimento de Plataformas de Governo para Geração e Divulgação de Informações e de Conhecimento. Artigo apresentado em cumprimento a requisito parcial de concurso para professor no INE/UFSC. Florianópolis, 2003.

LATTES. Conheça o Lattes. Disponível em: < http://lattes.cnpq.br/conheca/con_func.htm>. Acesso em : 11 out 2005.

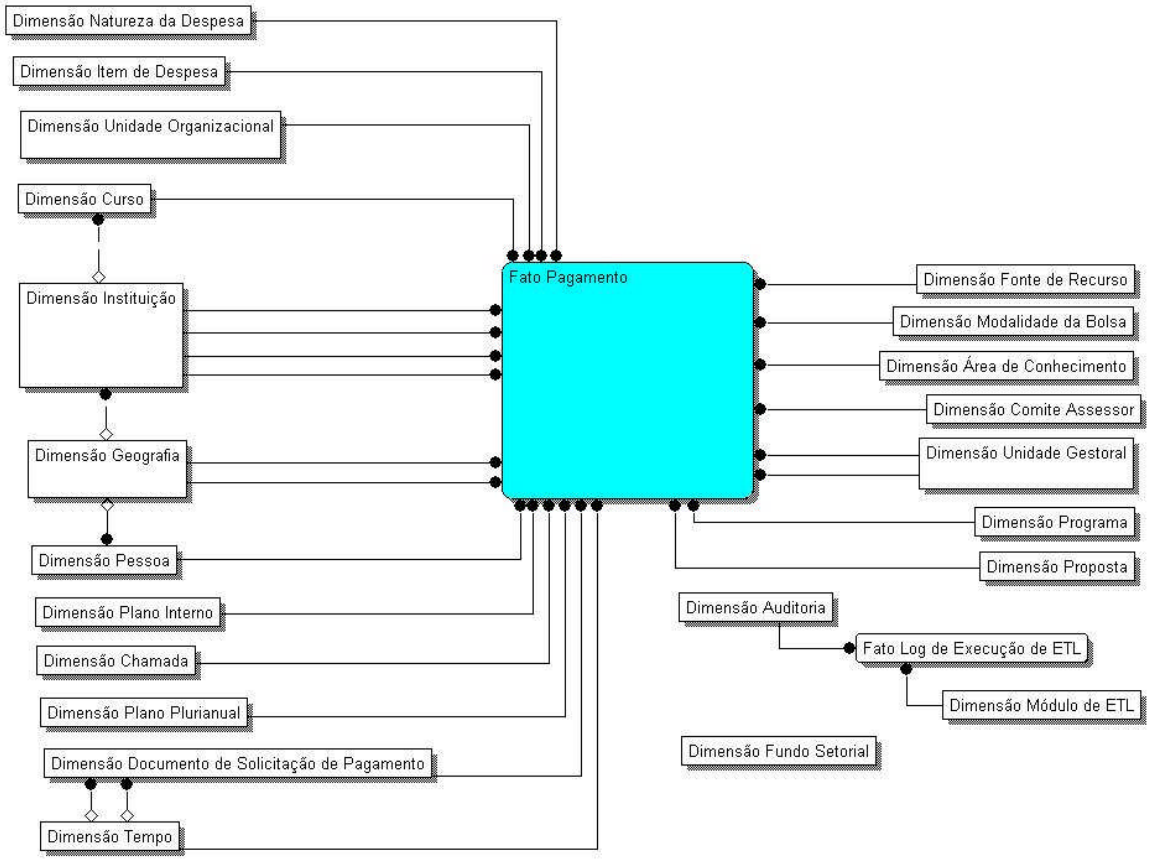
SINGH, Harry S. *Data Warehouse*: Conceitos, Tecnologias, Implementação e Gerenciamento. São Paulo: Makron Books, 2001.

SINHA, A.P.; SEN, A. A Comparison of Data Warehousing Methodologies. Communications if the ACM, Vol. 48, No. 3. Março de 2005.

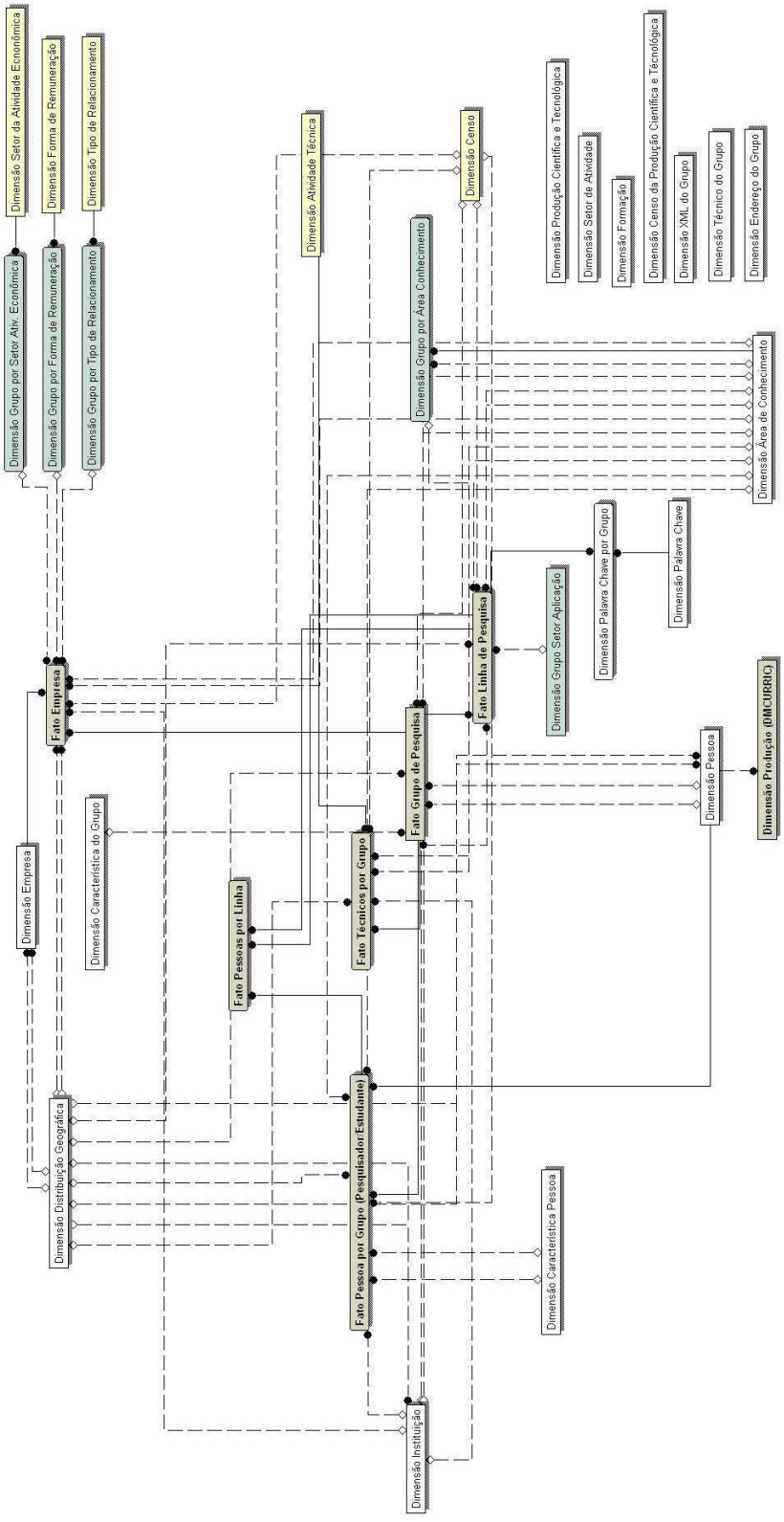
STUDER, R. et al. Situation and Prospective of Knowledge Engineering. In: CUENA, J. et al. (Ed.). Knowledge Engineering and Agent Technology. IOS Series on Frontiers in Artificial Intelligence and Applications. IOS Press, 2000. Available: <http://wwwdb. stanford.edu/%7Estefan/paper/2000/ios_2000.pdf>.

TISSOT, H.C., Proposta para Documentação de requisitos em projetos de Data Warehouse. Florianópolis, 2004. Dissertação (Mestrado em Engenharia de Produção) - Engenharia de Produção e Sistemas, UFSC.

Anexo I – Modelo do *DM* de Fomento



Anexo II – Modelo do *DM* de Grupos de Pesquisa



Anexo III – Modelo do *DM* de Currículos